DOCUMENT RESUME

ED 277 959 CG 019 641

AUTHOR	Kirby, Douglas
TITLE	Sexuality Education: A Handbook for the Evaluation of Programs.
INSTITUTION	Mathtech, Inc., Arlington, VA.
SPONS AGENCY	Center for Population Options, Washington, DC.; Centers for Disease Control (DHHS/PHS), Atlanta, GA.
PUB DATE	84
NOTE	179p.; For the complete report, see CG 019 636-642.
AVAILABLE FROM	Network Publications, 1700 Mission St., Suite 203, P.O. Box 1830, Santa Cruz, CA 95061-1830.
PUB TYPE	Guides - Non-Classroom Use (055) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Adolescents; Elementary Secondary Education; *Evaluation Methods; Parent Child Relationship; *Program Evaluation; *Research Methodology; *Sex Education; *Sexuality

ABSTRACT

This document is the fifth volume of a six-volume report on sexuality education. This volume is based on the methods used and the experiences encountered in the evaluation of the nine exemplary sexuality education programs contained in the first volume of the report. The present volume discusses the need for evaluation of sexuality education programs; selection of program characteristics and outcomes to be measured; experimental designs; survey methods; questionnaire design; and procedures for administering questionnaires, analyzing data, and using existing data. The volume focuses primarily upon the evaluation of sexuality education programs in the classroom for young people but also discusses the evaluation of peer education programs, one-day conferences, and programs for parents. The chapters discuss sequentially the important steps in conducting evaluation research. The appendix contains questionnaires and assessment inventories concerned with knowledge, attitudes and values, behavior, course evaluation, and course impact, and includes a course assessment questionnaire for parents. Thirteen figures and tables are included. (NB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED277959

CG 019641

SEXUALITY EDUCATION:

A Handbook for the Evaluation of Programs

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Steven Bignell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by Douglas Kirby

BEST COPY AVAILABLE

Sexuality Education:

A Handbook for the Evaluation of Programs

Developed at Mathtech, Inc.
by Douglas Kirby, PhD

Network Publications, Santa Cruz, 1984

Final report to the U.S. Department of Health and Human Services,
Public Health Service, Centers for Disease Control,
Center for Health Promotion and Education.

The opinions expressed in this report are those of the author(s) and do not necessarily reflect those of the U.S. Government.

Developed by Mathtech, Inc., 1401 Wilson Blvd., Suite 930,
Arlington, VA 22209
Telephone: (703) 243-2210

Part of this research was supported by the Center for Population Options,
2031 Florida Ave., N.W., Washington, D.C. 20009
Telephone (202) 387-5091

For ordering information contact:
Network Publications, a division of ETR Associates
P.O. Box 8506
Santa Cruz, CA 95061-8506
Telephone: (408) 429-8922

CONTENTS

ACKNOWLEDGMENTS ix

PREFACE xi

Background of This Project
Overview of This Report

INTRODUCTION 1

The Need for Evaluating Sexuality Education Programs
An Appraisal of Sex Education Evaluation
About This Volume

PART I: DEFINING THE PROBLEM

1 DEFINING THE BASIC PARAMETERS OF THE STUDY 5

Initial Decisions about the Goals and Resources of the Evaluation
Basic Decisions about Methodological Approaches
The Role of Values in Conducting Research
Overall Approach to Designing and Evaluating Programs

2 IDENTIFYING IMPORTANT FEATURES, OUTCOMES, AND GOALS OF PROGRAMS 15

Task 1: Establish Major Goals
Task 2: Specify Behavioral Objectives Leading to These Goals
Task 3: For Each Behavioral Objective, Specify the Necessary Knowledge, Attitudes, and Skills
Steps Within Each Task
Task 4: Identify Unexpected or Undesired Effects

PART II: SELECTING THE OVERALL DESIGN

3 USING EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS 21

One-Shot Case Study
One-Group Pretest Posttest Design
Nonequivalent Pretest Posttest Control Group Design
Randomized Pretest Posttest Control Group Design
Delayed Treatment Design
Pretest and Multiple Posttest Control Group Design
Posttest-Only Control Group Design
Alternative Tests Design
Solomon Four Group Design
Time Series Design
Summary

4 CONDUCTING SURVEYS 33

Types of Surveys
Surveys Versus Experimental Designs
Uses in Designing and Evaluating Programs
Conclusion

PART III: DESIGNING QUESTIONNAIRES

5 EMPLOYING THE FUNDAMENTALS OF QUESTIONNAIRE DESIGN 39

Important Steps in Designing Questionnaires
Determining the Important Features and Outcomes That Should be Measured
Constructing the Questionnaire
Pretesting the Questionnaire
Assessing Reliability
Assessing Validity
Conclusion

6 MEASURING PARTICIPANTS' ASSESSMENTS OF THE PROGRAM 51

Using Participants' Assessments
Writing Questions
Choosing Response Categories

7 DESIGNING KNOWLEDGE TESTS 55

Using Existing Knowledge Tests
Selecting Formats
Selecting the Number of Questions in Each Content Area
Writing Questions
Conducting an Item Analysis
References

8 DESIGNING QUESTIONNAIRES TO MEASURE ATTITUDES, VALUES, AND FEELINGS 63

Selecting Important Attitudes and Values to Measure
Using Scales Constructed by Others
Selecting the Best Scales
Constructing and Pretesting the Scales
References

9 DESIGNING QUESTIONNAIRES TO MEASURE BEHAVIOR AND SKILLS 73

Determining the Important Behaviors to be Measured
Constructing the Questionnaire
Pretesting the Questionnaire
Assessing Reliability and Validity

PART IV: ADMINISTERING THE STUDY

10 SELECTING A SAMPLE 81

Selecting a Sample Size
Improving the Randomness of a Sample
Improving the Response Rates
The Sampling of Programs
Reference

11 ADMINISTERING QUESTIONNAIRES 87

Obtaining Approval
Selecting a Test Administrator
Selecting Dates
Ensuring Voluntariness While Encouraging Cooperation
Ensuring Anonymity
Using Identification Numbers
Giving Directions and Answering Questions
Allowing Sufficient Time

12 USING UNOBTRUSIVE MEASURES 93

Using Unobtrusive Measures in Other Fields
Using Unobtrusive Measures to Measure Contraceptive, Pregnancy, and STD Rates
Using Unobtrusive Measures to Evaluate Other Goals
Reference

PART V: ANALYZING THE DATA

13 PREPARING DATA FOR ANALYSIS 99

Doing the Analysis by Hand Versus Computer
Coding Questionnaire Data
Keypunching the Data
Setting up Keypunched Data on the Computer
Creating an SPSS or SAS Program File
Cleaning the Data
Reference

14 STATISTICAL ANALYSIS 105

Kinds of Data
Descriptive Statistics
Inferential Statistics
Meaningfulness of Results
Recommended Statistics Books

15 WRITING THE EVALUATION REPORT 119

Planning the Writing Project
Presenting Quantitative Results
Presenting Nonquantitative Data
Dilemmas in Writing and Publishing the Results
Suggested Readings

PART VI: EVALUATING SPECIFIC KINDS OF PROGRAMS

16 EVALUATING SPECIFIC KINDS OF PROGRAMS 127

Comprehensive Programs Lasting About a Semester
Short Structured Courses Lasting 1 or 2 Weeks
One-day Conferences
Peer Education Programs
Parent/Child Programs
Conclusions

APPENDIX

KNOWLEDGE QUESTIONNAIRE 137

ATTITUDE AND VALUE INVENTORY 145

Scales in the Attitude and Value Inventory 150

BEHAVIOR INVENTORY 153

KNOWLEDGE, ATTITUDE, AND BEHAVIOR QUESTIONNAIRE 161

COURSE EVALUATION 169

ASSESSMENT OF COURSE IMPACT 173

COURSE ASSESSMENT FOR PARENTS 177

FIGURES AND TABLES

FIGURE 1-1 Major Stages in Designing and Implementing a Program and Evaluation 13

TABLE 6-1 Examples of Different Response Categories 54

TABLE 8-1 A Likert Scale to Measure Self Esteem 65

TABLE 8-2 Semantic Differential Scale Used to Measure Attitude toward Contraception 68

FIGURE 14-1 Knowledge Test Scores Presented as Original Raw Data 108

FIGURE 14-2 Pretest and Posttest Scores Ordered in Arrays 109

FIGURE 14-3 Knowledge Test Scores Presented in Frequency Distributions . 110

FIGURE 14-4 Knowledge Test Scores Presented as Percentage and Cumulative Percentage Distributions 111

FIGURE 14-5 Knowledge Test Scores Presented in Grouped Frequency Distributions 112

TABLE 15-1 Mean Pretest and Posttest Scores on a 40-Item Multiple Choice Test 121

TABLE 15-2 Mean Pretest and Posttest Scores on a 40-Item Multiple Choice Test for a Sexuality Education Class and Its Control Group 121

FIGURE 15-1 Number of People Receiving Different Scores on Pretests and Posttests 123

FIGURE 15-2 The Mean Knowledge Test Scores for Students on the Pretest and Posttest 123

ACKNOWLEDGMENTS

Walter J. Gunn, Ph.D., Director, Research and Evaluation, developed the approach for this entire project, initiated the contract, monitored progress, and provided technical assistance, guidance, administration, and support. Clearly, this and the other volumes would not have been possible without his continuing effort and support.

Judith Alter, Jesse Blatt, Nancie Connolly, Lynne Cooper, Bernard Kirby, Guy Parcel, and Peter Scales carefully read the volume and made numerous helpful comments. Lynne Cooper made numerous substantive suggestions that were especially helpful and improved the volume.

Ann Thompson Cook spent many hours editing this volume. She has made it much more clear, concise, and readable. Karen Allan provided considerable help with the typing and production.

PREFACE

Background of This Project

During the mid 1970's the Carter administration recognized the large number of unintended teenage pregnancies in America and sought solutions to this major problem. That administration recognized that one potentially effective solution was sexuality education. Consequently, it asked the Center for Health Promotion and Education (formerly the Bureau of Health Education) in the Centers for Disease Control to identify, improve, and evaluate promising approaches to sexuality education.

The current project followed an earlier 1978 contract that the Center for Health Promotion and Education awarded to Mathtech to identify promising programs and to develop evaluation methods. In that project, Mathtech, with the help of many sexuality educators and other related professionals:

- identified and rated about 200 features and outcomes of programs potentially important to reducing pregnancy and increasing psychological health
- reviewed the literature on the effects of sex education programs
- identified 10 promising programs representing several different approaches
- developed questionnaires and other methods to more effectively measure the important outcomes of these promising programs
- summarized the work in a six-volume report entitled An Analysis of U. S. Sex Education Programs and Evaluation Methods.

In 1979 the Center awarded Mathtech a second contract to help improve and then evaluate 10 of the promising sexuality education programs. Mathtech selected 10 exemplary programs that represented a variety of different approaches to sexuality education. The programs include 6-hour programs, semester programs, conferences, programs for young people alone and for young people and their parents together, peer education programs, both school and non-school programs, and both educational and clinic approaches. Mathtech:

- conducted an initial evaluation of each program
- suggested numerous changes which the sites incorporated
- offered training to the program staffs
- provided some materials and other kinds of support
- then carefully evaluated the programs.

The results of this contract are summarized in this report.

The Organization of This Report

The complete report contains several separate volumes and an Executive Summary which summarizes the first volume. Although all of the volumes are an integrated package which we hope will meet many varied needs of educators, evaluators, and policy makers, some of the volumes will have particular interest for selected groups of people, and each volume is complete and can be used independently of the others.

Sexuality Education: An Evaluation of Programs and Their Effects...An Executive Summary summarizes first the existing information on sexuality education in the United States and then the overall design, methods, and major findings of this evaluation.

The first volume, Sexuality Education: An Evaluation of Programs and Their Effects, summarizes the structure and content of sexuality education in the United States, reviews the literature on the effects of sexuality education, describes the evaluation methods, provides a description of and the evaluation data for each program, and summarizes the effectiveness of different approaches in meeting different goals.

The second volume, Sexuality Education: A Guide to Developing and Implementing Programs, provides suggestions for developing and implementing effective educational and clinic-based approaches to sexuality education. It discusses the reasons for and nature of responsible sexuality education and describes approaches to building a community-based program, selecting teachers and finding training, assessing needs of the target population, and designing and implementing programs for them. It also provides suggestions for evaluating programs.

The third volume, Sexuality Education: A Curriculum for Adolescents, is based upon the curricula of the most comprehensive programs. These programs increased knowledge and helped clarify values. The curriculum consists of the following units: Introduction to Sexuality, Communication Skills, Anatomy and Physiology, Values, Self Esteem, Decisionmaking, Adolescent Relationships, Adolescent Pregnancy and Parenting, Pregnancy Prevention, Sexually Transmitted Diseases, and Review and Evaluation. Each unit contains a statement of goals and objectives, an overview of the unit contents, several activities that address the goals and objectives, and wherever needed, lecture notes and handouts.

The fourth volume, Sexuality Education: A Curriculum for Parent/Child Programs, is based upon the parent/child program which increased knowledge and parent/child communication. The curriculum includes several suggested course outlines and the following units: Introduction to Course; Anatomy, Physiology, and Maturation; Gender Roles; Sexually Transmitted Diseases; Reproduction; Adolescent Sexuality; Birth Control; Parenting; and Review. Each unit contains several activities and, wherever necessary, lecture notes and handouts.

This fifth volume, Sexuality Education: A Handbook for Evaluating Programs, is based upon the methods we used and our experiences in evaluating these programs. It discusses the need for evaluation of sexuality education programs; selection of program characteristics and outcomes to be measured; experimental designs; survey methods; questionnaire design; and procedures for administering questionnaires, analyzing data, and using existing data.

A sixth volume, Sexuality Education: An Annotated Guide for Resource Materials, reviews books, films, filmstrips, curricula, charts, models, and games for youth in elementary school through high school. For each resource, the guide lists the distributor, length, cost, and recommended grade level, and provides a discussion of the material. This volume differs from the others in that it was not funded by the government and is not part of the final report. However, it will be useful to people developing programs.

INTRODUCTION

The word "evaluation" commonly refers to a variety of informal and formal, nonsystematic and systematic assessments and judgments. This guide will use "evaluation" in its more narrow and scientific sense, as the formal and systematic process of collecting information about a program in order to determine the effectiveness of that program and to make better decisions about that program.

There are several different models for collecting information. This guide uses primarily a goal-attainment model which generally has four major steps: 1) defining the measurable program goals and objectives, 2) designing methods of measuring and quantifying those goals and objectives, 3) collecting data that measure them, and 4) reaching conclusions about the extent to which the goals and objectives are reached.

The goal-attainment model can be contrasted with a goal-free model, which involves measuring all outcomes, not just goals, and a systems model, which involves analyzing costs and benefits.

The Need for Evaluating Sexuality Education Programs

As a general rule, any social activity, program, or policy designed to alleviate social problems should be carefully evaluated whenever 1) the program or practice and decisions about it are important, 2) the outcomes cannot be assessed without an evaluation, and 3) informal and nonsystematic observations and information cannot provide sufficient data for decisionmaking. Even when informal observations are sufficient for decisionmaking, programs should be systematically evaluated whenever other people require evidence about the success of the program. Without careful evaluation, ineffective practices or programs may be maintained, and effective practices or programs may be canceled.

The evaluation of sexuality education programs is particularly important. First, sexuality education programs are designed and implemented to improve the lives of young people in very important ways. For example, some educators hope sexuality education will improve interpersonal communication, decisionmaking, responsibility, social relationships, and self esteem, and that it will reduce unwanted sexual activity, unprotected intercourse, unwanted pregnancies, sexually transmitted diseases, rape, and some sexual dysfunctions. To achieve such outcomes, substantial and increasing amounts of time, money, and other resources have been devoted to sexuality education.

Second, nonsystematic observations and past research have left unanswered many important questions about sexuality education programs.

- What are the long term effects of sexuality education?
- How does it affect students' attitudes and behaviors?
- Does it reduce unwanted pregnancy and sexually transmitted diseases?
- Does it improve young people's communication with parents?

- Are shorter programs more cost effective than semester programs or vice versa?
- Are separate courses more effective than units which are part of other courses (e.g., a sexuality unit within a science course)?
- What topics are most important?
- What characteristics of teachers are most important?
- What kinds of activities -- lectures, discussions, role-playing, films -- are most effective?

These and other questions remain unanswered. Moreover, when these questions are answered about sexuality education in general, similar questions about individual programs with specific structures (e.g., number of sessions and length), specific personnel, and specific materials will still remain. Effectiveness should be determined in each case.

Third, sexuality education has been the subject of heated controversy throughout the country. Opponents claim that sexuality education has many negative effects; proponents claim that it has many positive effects. Neither side can prove its case. Well validated research can eventually help calm these conflicts.

Fourth, program providers often need the clarity and realism that evaluation produces. Too many programs have vague and unrealistic goals. As staff members anticipate the evaluations, or become involved in them, they become both more clear and more realistic about the program and its goals. Thus, the mere process of evaluating programs can help improve the programs.

Both the importance of sexuality education and the need for evaluation are demonstrated by the many people who are currently asking important and difficult questions about sexuality education. Each month reporters from newspapers, magazines, and radio and television stations request information on the amount of sexuality education in schools, the comprehensiveness of programs, and the effects of programs. Each month several Congressional representatives request information about the effects of sexuality education programs. They ask whether programs reduce unwanted pregnancies, increase self esteem, and improve the psychological health of adolescents. Each month educators ask about the evidence for the success of programs and the realism of meeting expected goals. Unfortunately, most of these questions and requests for information cannot be adequately answered because the necessary research has not been conducted or completed.

The importance of evaluating sexuality education programs is further exemplified by the surprising results of evaluating programs in other fields. For example, many states, observing that teenagers were involved in a disproportionate number of automobile accidents, developed drivers′ education programs for them. The educators and others believed that such programs would increase the students′ knowledge about safety, make them more responsible, increase their driving skills, and consequently reduce their number of accidents. However, several recent studies demonstrate that drivers′ education helps teenagers drive at an earlier age and, therefore, it may ultimately increase the number of accidents and deaths among teenagers. These results are just the opposite of expectations.

You should not conclude from the example above that if drivers′ education increases accidents, then sexuality education will increase sexual activity and pregnancies. There is a critical difference between the two programs -- drivers′ education is designed to teach students to drive, but sexuality education is NOT designed to teach young people to have sex. However, the example does demonstrate the importance of making sure that our important social and educational programs, including sexuality education, are having the effect(s) we intend.

An Appraisal of Sex Education Evaluation

Many evaluations of sexuality education programs have not been true evaluations -- they have described the programs, but have not carefully evaluated the effects of the programs.

Of those evaluations that have examined the effects of the program, most have employed some type of experimental or quasi-experimental design. In such studies, the sexuality education class is considered the experimental group, and occasionally some other class is treated as the control group. Evaluators then give questionnaires both before and after the course to both the experimental and control subjects. This kind of design can provide good evidence for the effects of the course.

Unfortunately, there are numerous limitations with the evaluations that have employed this design:

- Many studies have evaluated single programs which may or may not be representative of all sexuality education programs, and thus it is difficult to generalize from them to other courses.
- Because evaluators have rarely been able to randomly assign students to experimental and control groups, some self-selection factors may have affected their results.
- Very few evaluations have measured effects beyond the end of the program.
- Most questionnaires focused upon knowledge and failed to measure many important attitudes and behaviors.
- Many questionnaires have been poorly designed.
- Many evaluations reported the statistical significance of the change in students, but few evaluations reported the magnitude of the change and its theoretical or practical significance.

Fortunately, during the last few years, an increasing number of people have been recognizing the need for evaluation, and there has been considerable growth in the evaluation of sexuality education. Research groups are developing and disseminating new evaluation materials; professional organizations are offering special sessions or seminars on evaluation; a few research groups such as E.T.R. Associates, Mathtech, and Johns Hopkins University are conducting more formal evaluations of programs, and an increasing number of schools, clinics, and other youth-serving organizations are actually evaluating their programs. Thus, the direction is positive, but the need will be met only with considerable effort for many years.

About This Volume

This guide introduces methods of evaluating sexuality education programs. It discusses the need for evaluation; selection of program characteristics and outcomes to be measured; experimental designs; survey methods; questionnaire design; and procedures for administering questionnaires, analyzing data, and using existing data. It provides both fundamental principles and practical suggestions for evaluation. In the appendix are reliable, valid questionnaires that have been used in the evaluation of sexuality education programs.

The volume focuses primarily upon the evaluation of sexuality education programs in the classroom for young people but also discusses the evaluation of peer education programs, one-day conferences, and programs for parents. Educators can apply the same principles and methods to other kinds of programs.

Much of the volume is written in sufficient detail for the lay person with only a beginning knowledge of evaluation methods, but contains numerous practical suggestions and several sections that should be helpful to the more advanced methodologist as well.

The chapters in this guide discuss sequentially the important steps in conducting evaluation research. Many aspects of a good design are interdependent, however, so that one must continually think ahead to subsequent steps when making decisions about earlier steps.

If you have never conducted an evaluation, this volume, and evaluation more generally, may appear intimidating. If so, start small -- peruse this volume and conduct a small and relatively simple evaluation. Use a simple design, measure only a few outcomes, and use a small sample. Then reread this volume and improve your design and questionnaires. Your initial evaluation may provide useful information for your program and will help you learn about evaluation so that your subsequent evaluation efforts can be more rigorous and valid.

If you have conducted several evaluations, you may already be familiar with parts of this volume. Feel free to skim those parts and to focus on those parts that are most informative.

CHAPTER 1

DEFINING THE BASIC PARAMATERS OF THE STUDY

Evaluation is a process of systematically collecting information so that people can make better decisions about programs. However, people in different positions need varying kinds of information to make their decisions. For example, staff and administrators may want information about the impact of the program while funders may prefer information on the numbers of people served and the costs of the program. Collecting these different kinds of information requires different methods. The different needs of different groups are legitimate, but you will need to establish priorities. Before you begin to design specific tools for collecting data, you should make a number of basic decisions about the overall goals, scope, and structure of the evaluation.

Initial Decisions about the Goals and Resources of the Evaluation

Who Will Use the Evaluation Results?

For any evaluation of sexuality education programs, there are several possible users:

- The educators or instructors themselves
- The administrators of the program
- The people or agencies funding the program
- Other professional educators or organizations involved with sexuality education and interested in putting on a program
- Lay members of the community interested in the program
- Groups served by the program.

Each of these groups may have legitimate, but different needs. Often their needs will overlap so that any evaluation will be helpful to many of them. However, you will have to make many decisions which will make the evaluation better suited to one group than another. Select the primary users and then direct the evaluation to them.

What Is the Purpose of the Evaluation?

The different groups mentioned above may have different purposes for the evaluation. Moreover, any one group may also have more than one purpose. Users may want the evaluation:

- to describe the contents of the existing program
- to assess the impact of the program upon the participants
- to assess the relative effectiveness of different program components
- to estimate the number of people being served

- to assess the total cost of the program and the cost per person served and/or
- to identify ways to improve the program.

Once again, you will need to decide which of these different purposes take priority for the evaluation.

Which Program Components Will Be Evaluated?

Obviously, you need to know what programs or components you are going to evaluate before you can evaluate them. In schools, the sexuality education program is usually well defined. Typically it includes the sexuality units or courses in the school curriculum and the instruction is given to classes of students which meet at specified times in the classroom. However, nonschool programs may be less clearly delineated or may have many components. For example, a youth program may include regular group discussions at the agency, occasional parent/child activities, occasional films for people who drop in, outreach efforts at health fairs, and media public information spots. These different components may be linked in different ways, with some people participating in two or more components and others participating in only one.

If a program has multiple components, you may want to carefully define each component, decide which components to evaluate, and then measure the unique contribution of each component. Alternatively, you may want to measure the cumulative and interactive effect of all the components. This will give you evidence for the success of your entire program, but will not help you judge the relative importance of each component.

Is the Study Feasible?

When choosing the basic parameters of the study -- the kind of experimental design or survey, the basic kinds and numbers of questions to be asked, the approximate number of people in the sample -- you need to constantly keep in mind the resources available to your study. You must be able to answer the following questions affirmatively:

- Are the goals of the evaluation realistic?
- Are the funds, labor, and other necessary resources available to complete the evaluation?
- Is the proposed evaluation politically feasible? Will the necessary groups support it? Can the evaluators deal effectively with any opposition groups that may try to block it?
- Is it possible to obtain the desired data? Will the proposed respondents to the questionnaires actually complete the questionnaires? Will they find the questions acceptable and not too sensitive or personal? Can they complete the questionnaires in a reasonable period of time? Will their answers be reliable and valid?
- Are the resources available for data analysis and report writing?
- Can the report be disseminated appropriately?

Two resources are particularly important: the willingness of people to participate in and help with the evaluation and the availability of funds for materials, computer time, and professional help. The willingness of participants is perhaps the most critical. If program participants are not willing to fully cooperate with the evaluation, then its validity will be seriously compromised. If there are relatively few funds, you may still be able to complete a valid evaluation, but doing so will be more difficult. Occasionally, you can obtain a small grant to support your research. Otherwise, you can reduce costs without sacrificing quality by:

- collecting only the most important information that you need
- modifying and then using previously validated questionnaires instead of creating questionnaires from scratch
- collecting data from only a sample of people instead of the entire population
- having program participants and staff collate questionnaires, put them in envelopes, and later code them
- scoring questionnaires by hand, instead of using computer facilities or using tests which can be machine scored
- obtaining statistical advice and analytic support from graduate students who might wish to use the evaluation as part their work toward an advanced degree.

Who Will Conduct the Evaluation?

Increasingly a wide variety of people are conducting evaluations of programs. They range from teachers or clinic counselors with relatively little experience in evaluation to professional methodologists with substantial experience in evaluation. Ideally, the person conducting an evaluation would have:

- familiarity with basic methodological concepts such as experimental designs
- previous experience conducting evaluations
- skills in coordinating and administering
- freedom and ability to conduct an unbiased evaluation.

However, if you do not have previous experience, you can still collect information with the use of materials like this handbook and some advice from consultants. As suggested in the Introduction, you can first conduct a relatively simple evaluation, and as you become more experienced, then improve the design, questionnaires, and validity of your evaluation.

Basic Decisions about Methodological Approaches

Descriptive Versus Evaluative Information

In evaluation, there is an important distinction between descriptive information and evaluative information. The former simply describes the existing program: its length, the number of hours the classes meet, the topics covered, the different kinds of activities used, the characteristics of the staff, the number of students that attend. Evaluative information describes the quality and success of the program by measuring its effects.

Too often people have evaluated their programs by presenting only descriptive information about their programs. Unfortunately, the existence of excellent resources does not always assure desired effects. Therefore, if you want to measure the actual effectiveness of your program -- whether or not it has a desired impact -- you should collect evaluative information and actually measure the effects.

This handbook focuses upon methods for obtaining evaluative information. Thus, it will not discuss methods for describing the program components, counting the numbers of people served, or estimating the costs of the program, but it will discuss methods for helping improve the program and methods for assessing the impact of the program.

Formative Versus Summative Evaluation

When collecting evaluative information (as opposed to descriptive information), there are two kinds of evaluation: formative and summative. In a formative evaluation, evaluators might lead group discussions or administer questionnaires in which they ask the program participants how they liked the program, what parts of the program they would change, and how they would improve it. Because the focus of formative evaluations is to give feedback as quickly as possible, such evaluations are commonly conducted during the course as well as at the end of the course. Educators who are primarily interested in improving their programs should conduct a formative evaluation. Such an evaluation is designed to provide more quickly the kinds of data educators need to improve their programs.

When evaluators are primarily interested in measuring the effectiveness of a program, they should conduct a summative evaluation. Such an evaluation focuses more directly upon the actual outcomes of the program and will thereby help other educators and policymakers decide whether they wish to adopt this program or other programs. Although a summative evaluation can help educators improve their program, a summative evaluation provides less direct information about specific ways to improve the program.

Because a summative evaluation reports the success of an entire program, and because that report may affect decisions about the continuation of that program or the adoption of that program elsewhere, that evaluation is especially important, and the methods must be especially valid and defensible. Thus, summative are more likely than formative evaluations to use experimental or quasi-experimental designs.

Experimental and Quasi-experimental Designs Versus Survey Methods

Experimental and quasi-experimental designs and surveys often involve the administration of questionnaires one or more times and may measure the effects of sexuality education programs. How, then, do they differ?

True experimental designs differ from quasi-experimental designs and surveys in one critical respect: experimental designs include the random assignment of people to the experimental and control groups. In the evaluation of sexuality education, the experimental group participates in the sexuality education program and the control group does not. The random assignment of people to the experimental and control groups will cause the two groups to be similar before the experimental group participates in the program. Then, if the two groups are different after the program, you may be able to attribute this difference to participation in the program. Some experimental designs also include the administration of questionnaires before the program (pretests) and after the program (posttests), and thereby allow you to actually measure change. Thus, experimental designs provide the best evidence for the causal impact of sexuality education programs.

Quasi-experimental designs do not include the random assignment of people to the experimental and control groups. However, they do have some of the other features of experimental designs. For example, they may include experimental and control groups even though people were not randomly assigned to them, and they may also include pretests and posttests. The evidence they provide for the causal impact of programs is poorer than that of true experimental designs, but better than that of surveys.

In contrast, surveys typically do not include any control over the participants. Respondents in the sample may have participated in no sexuality education program or in a variety of different sexuality education programs. Moreover, they may have participated in a program recently or long ago. Thus, they provide the poorest evidence for the impact of programs. They nevertheless can be useful in obtaining additional information about programs.

Experimental designs and survey methods are discussed fully in Chapters 3 and 4 respectively.

Normative Referenced Versus Criterion Referenced Methods

When teachers or researchers evaluate individuals or groups, they often give each person a score that is determined by that individual's performance relative to the performances of others. For example, some teachers give the top 10% of the students in a class an "A," the next 30% a "B." Similarly, people may be given percentile rankings on Graduate Record Exams or other tests of knowledge or skill. Such scores are based upon norms established by the group, and accordingly are called normative referenced measures. Such scores order people; they indicate that one particular person is more or less capable than others but do not tell anything about the absolute capability of that individual. Consequently, they are very useful whenever the object is to compare individuals. For example, graduate schools prefer normative referenced scores to select graduate students.

On other occasions, evaluators assign scores based upon some set of specified standards and not upon the relative performances of individuals. For example, people who take a driving test are given a score (pass or fail) that depends not upon their performance relative to others, but upon their ability to perform a set of specified driving tasks. Because these scores are based upon specified criteria, they are called criterion referenced measures. Such measures are useful in determining those areas in which individuals have sufficient knowledge or skills and those areas in which they need to improve.

Both normative and criterion referenced questionnaires may resemble each other in outward appearance. However, different steps must be completed to develop them. For example, when developing normative referenced questionnaires to measure knowledge, you should specify the different areas of knowledge that you wish to measure and write questions for each area. Some of these questions should be easy and others difficult so that the more informed students are separated from the less informed. When developing criterion referenced questionnaires, you should specify very carefully the exact knowledge facts to be known, develop questions for those facts, and then specify criterion levels for the percentage of questions that people should get correct in order to be able to perform some desired activity. When designing this criterion referenced questionnaire, you would have less concern about selecting both easy and difficult questions.

In the past, normative referenced measures were more commonly used, because they were developed first, and also because they are so commonly used to grade students in schools and universities. In general, however, criterion referenced measures are preferable. When evaluating programs, you will probably want to know whether the participants are learning the specific facts that are needed, and you will probably want to know whether they need additional help in specific areas. Criterion referenced measures can better provide this information.

Note that you can often develop questionnaires as criterion referenced measures and use them in normative referenced measurement (to rate individuals relative to one another), but you cannot do the reverse satisfactorily. That is, you cannot develop questionnaires with normative referenced measures and then use them in criterion referenced measurement (to ascertain whether specified standards have been met), because the specific content areas and the needed levels of competence will not have been specified. Thus, developing criterion referenced measures also provides more versatility.

One Versus Two or More Methods of Collecting Data

An important principle in methodology is that you should use at least two maximally different methods to collect evidence for the success of programs and then compare the conclusions that logically follow from each method. This is important because every method of evaluation has some assumptions, some biases, and some inevitable sources of error. Maximally different methods are less likely than similar methods to have the same assumptions, the same biases, and the same sources of error. Therefore, if two maximally different methods produce consistent conclusions, those conclusions are less likely to be caused by the same underlying assumptions, biases, and errors, and are more likely to be valid. If, on the other hand, conclusions derived from one method are inconsistent with those derived from a second method, then you know that either one or both of the methods contains some source of error. Being able to check the conclusions from one method against the conclusions from a second method can substantially increase the validity and the credibility of your conclusions.

Multiple sources of information. Most commonly evaluators of sex education programs get their most valid and complete information directly from the participants in the program. However, many times you can also obtain valuable information from the parents or teacher of the participants. Occasionally, you can get valid information from the school nurse and principal (if it is a school program), from peers, and from outside observors.

Multiple methods of information. There are numerous different methods of collecting data. These include questionnaires, unobtrusive measures (including extant data), direct observations, and group discussions.

Questionnaires can often provide the best data about programs. They have several advantages: they can be anonymous; they are systematic; they can cover a wide variety of topics; and they can include questions about activities outside of the classroom. However, they have two major disadvantages: they are self-reports and they require the cooperation of the students and possibly teachers and others. Thus, if respondents either intentionally or unintentionally answer questions incorrectly, the resulting data are invalid. Questionnaires are discussed in detail in Chapters 5 through 9.

Unobtrusive measures involve the collection of data without the knowledge or participation of the respondents. Thus, they overcome some of the problems of questionnaires. They include important kinds of data (e.g., pregnancy or STD rates)

23

that may have been collected by others for other reasons. Unobtrusive measures are discussed in Chapter 12.

Parents, teachers, principals, outside observors, and others can directly observe changes in the participants. Teachers, especially, can observe changes in the students' knowledge, attitudes, and comfort as they are expressed in classroom discussions, questions, and comments. Sometimes teachers talk with individual students after class and help them solve their problems, and in these ways learn about the impact that the course has had. All of these observations can be valuable, but they may not always be valid. Too often, teachers or program staff view the events in the program selectively; they focus more upon those events that indicate the program is effective but fail to fully consider evidence that indicates that the program is ineffective. Even if the staff's observations are accurate and unbiased, they will not be viewed as solid evidence for the success of a program. Thus, you should consider their observations a valuable source of insights, but you should not use them to reach final conclusions about the success of a program in a summative evaluation.

Group discussions can be particularly helpful in formative evaluations, but they provide much less valid data for summative evaluations. That is, they are not anonymous or systematic; they rely upon verbally expressed self reports; and consequently they are not likely to provide valid information about the actual effects of the program upon them.

In sum, no single method is best; the most valid conclusions can be obtained from the judicious use of two or more maximally different methods. Frequently, if you have to use a single method, questionnaires will provide the best evidence, but if you have valid pregnancy or STD rates, they may better measure the programs' impact upon these rates.

The Role of Values in Conducting Research

The subject of values is an important and frequently discussed topic in sexuality education courses. Similarly, it is an important topic in sexuality education research and even more generally in most social science research. The basic questions are, "Should your basic values affect your research?" "If so, how?"

One answer commonly held by social scientists is simple in principle, but not necessarily simple to implement: Values can or should affect your choice of the problem you study, but once you have selected a topic of study, your values should not affect your analysis or conclusions.

Practically, this means that you should consider your values when you decide whether or not to evaluate sexuality education programs, and perhaps, you should even consider your values when you think about the magnitude of that effort. Thus, it is perfectly acceptable and probably even preferable to consider the need and the costs -- both economic and social -- of research in sexuality education.

Practically, this also means that once you have defined the scope of the research, your desire to demonstrate that a program's success should not bias your research. That is, you should be committed to an unbiased, valid study and to the advancement of knowledge, but not to the demonstration of a program's success or failure.

At each stage of the evaluation process, there are many overt and subtle ways that you can bias the results. For example, when designing questionnaires, you may be tempted to ask questions that will probably reflect well on the program and neglect to measure outcomes that will probably be negative. When administering the questionnaires, you may be tempted to stress the importance of getting particular kinds of findings and thereby encourage students to bias their answers and to give you answers that please you. When reporting the results, you may be tempted to stress only the positive results and fail to report non-findings or negative results. Thus, there are many opportunities for you to introduce your own biases, and you must continually guard against that. You should be careful to use established procedures for guarding against bias, and you might consider involving in your evaluation consultants, university professors, or others who can serve as your super-ego.

If you allow your values and biases to enter into your evaluation process, they will block your accurate understanding of sexuality education, reduce the faith that people have in research findings about sexuality education (and other topics), and partially destroy the integrity of evaluation methods.

Overall Approach to Designing and Evaluating Programs

Major stages in designing and implementing a program and evaluation are diagrammed in Figure 1-1.

Determining the desired features and outcomes of programs before designing either the program or the evaluation is very important. The specified features and outcomes should guide the design of both the program and the evaluation. Fit between the two helps ensure that the program and evaluation are striving to achieve and evaluate the same set of goals and objectives. Fit helps prevent people from designing a program with one set of goals in mind or with goals loosely defined, then claiming during the evaluation that additional goals are important even though the program was not designed to achieve those goals, and then learning that the program did not meet these additional goals and is not effective.

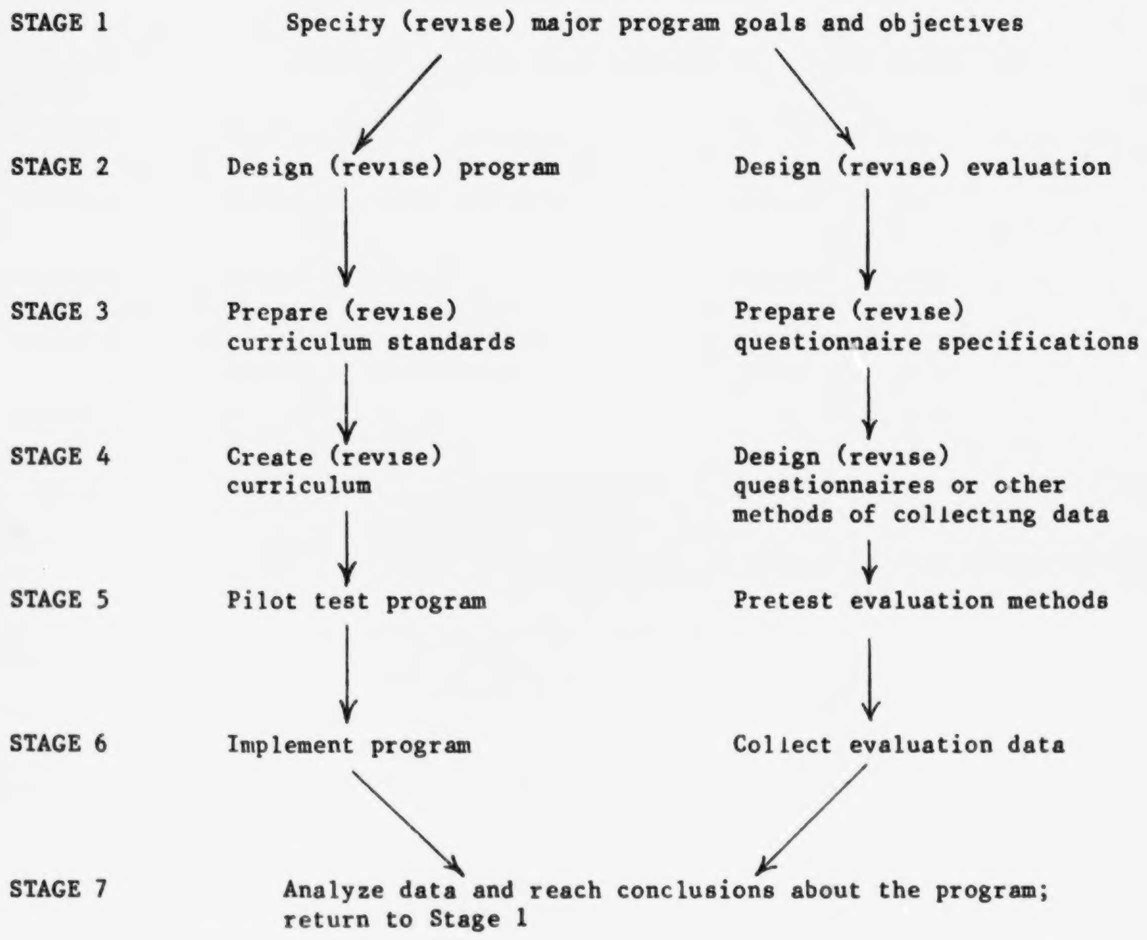
Of course, many people often design programs, implement them, observe problems, modify the program, and only then think about evaluating the program. Although designing the program and the evaluation methods simultaneously is better, it is possible to design valid evaluations well after programs have been implemented. In this case, you should follow the bottom set of tasks in Figure 1-1, and you must be sure that the stated goals really are the goals of the program.

After specifying the major goals and objectives of the program, you need to design the basic structure of both the program and the evaluation. For example, when designing the basic structure of the program, you should consider the needed length, the number of sessions, the participants, etc. When you design the basic structure of the evaluation, you should answer the questions discussed in this chapter about descriptive versus evaluative information, formative versus summative evaluation, experimental designs versus surveys, etc.

Next, you and your colleagues should specify more precisely the contents of the curriculum and the questionnaires. For example, you should specify the particular knowledge facts that should be covered, the specific attitudes that should be encouraged, and the specific skills that should be taught. The curriculum should include not only the specific topics that had been previously specified, but also the actual activities or processes that will teach the needed knowledge, attitudes,

Figure 1-1

Major Stages in Designing and Implementing a Program and Evaluation


STAGE 1
Specify (revise) major program goals and objectives
STAGE 2
Design (revise) program
Design (revise) evaluation
STAGE 3
Prepare (revise) curriculum standards
Prepare (revise) questionnaire specifications
STAGE 4
Create (revise) curriculum
Design (revise) questionnaires or other methods of collecting data
STAGE 5
Pilot test program
Pretest evaluation methods
STAGE 6
Implement program
Collect evaluation data
STAGE 7
Analyze data and reach conclusions about the program; return to Stage 1

and skills. Your questionnaires should also be as complete as possible, including the actual questions that you will pilot test.

The next step is to pilot test both the curriculum and the evaluation methods. This is a very important step, for it will probably suggest innumerable improvements that should be made. After improving both program and evaluation methods, you should implement and evaluate the actual program. The results of your analysis can then lead to suggested improvements in your program if the evaluation finds weaknesses or to possible expansion of your program if the evaluation is very positive.

This sequence can actually be considered a closed loop. Once you have recommendations for improvements or expansion, you should then revise and/or expand the program and reevaluate it. Evaluation, then, should be a continuing process of clarifying goals, designing or improving the program, evaluating the program, and improving it again.

CHAPTER 2

IDENTIFYING IMPORTANT FEATURES, OUTCOMES, AND GOALS OF PROGRAMS

Frequently people evaluate programs by simply describing a program and assessing the number of people that participate in various activities or components of the program. Although these descriptions are helpful, they do not contain the critical part of an evaluation, namely, the assessment of the effects of the program upon the participants. Especially in a summative evaluation, you should carefully measure the consequences of the program as well as describe the processes taking place in the program.

When evaluators do assess the effects of programs, they too frequently give insufficient thought to systematically determining their important goals and behavioral objectives of the program. For example, many assess the program's impact on knowledge, simply assuming that improved knowledge will lead to desired changes in attitudes, skills, decisionmaking, and behavior. Clearly, this assumption is not always true.

If the goals of your program either explicitly or implicitly include goals about changes in attitudes, skills, or behavior, then you should measure the change in all of these. Failure to do so can substantially reduce the effectiveness of both the program and the evaluation. If you fail to include an essential objective in your program when you design a program, then that program may be much less effective. Similarly, if you do not measure an important outcome of your program, then you are not fully evaluating your program. In sum, you should carefully specify and then evaluate the important goals and objectives of your program.

Task 1: Establish Major Goals

The major goals of the program can be rather broad and you can choose several of them. Program planners and educators have used a variety of strategies to create an initial list of goals. They have 1) written down every idea arising out of group "brainstorming"; 2) studied other existing lists of goals; and 3) observed the contents of different programs and reflected upon the goals of the activities.

Following are a list of some goals that have been adapted from Volume III of this report, Sexuality Education: A Curriculum Guide for Adolescents. These goals are only examples; you should consider these and others and create your own.

- Students will have a greater understanding of their own values, their families' values, and their culture's values and will behave more consistently with those values.

- Students will behave in ways that increase their own selt esteem and the self esteem of others.

- Students will use a systematic decisionmaking process to make important decisions about social and sexual behavior so that their behavior is consistent with their values and goals.
- Students will enhance their communication about sexuality and other topics with parents, peers, and significant others.
- Students will enhance interpersonal relationships.
- Students will avoid social and sexual activity that is unwanted or inconsistent with their values.
- Students will have fewer unwanted pregnancies.
- Students will reduce the risk of getting and spreading sexually transmitted diseases.

Task 2: Specify Behavioral Objectives Leading to These Goals

Consider each goal, one at a time, and specify the important behaviors that facilitate that goal. Behavioral objectives differ from goals primarily in that they are substantially more specific. Thus, you will frequently have several objectives for each goal, although sometimes you may have only one.

Each behavioral objective should be:

- clear; if it is not clear, then it may confuse both the educators and the evaluators.
- unidimensional; if the objective has more than one component, it should be broken up into more than one objective.
- equally specific; some should not be general and others very specific.
- reasonably achievable; otherwise there is little reason to entail the cost of trying to achieve and/or evaluate it.
- measurable; if you cannot measure an objective, you may still wish to include it as an objective for your program, but there is little reason to include it in the evaluation.

For example, given the goal above, "Students will have fewer unwanted pregnancies," two possible behavioral objectives are:

- Some students will avoid unintended pregnancy by abstaining from sexual intercourse.
- Students who are sexually active will avoid unintended pregnancy by using effective forms of birth control.

Task 3: For Each Behavioral Objective, Specify the Necessary Knowledge, Attitudes, and Skills

Consider each behavioral objective, and then specify all the knowledge areas, attitudes, and skills that are needed for each objective. (Hereafter, the knowledge areas, attitudes, and skills will be called simply the KAS components.) Often you will find many different KAS components for each objective. You can use these as the basis for both the program curriculum and the evaluation questionnaires.

Specifying the objectives is especially important in sexuality education evaluation. Sometimes you cannot directly measure the important behavioral goals and you can only measure the changes in knowledge, attitudes, and skills that experts believe will lead to the behavioral goals. In such situations, it is especially important to define precisely the knowledge components, attitudes, skills, and behaviors that are needed to reach the overall goals.

The KAS components should have the same qualities as the behavioral objectives. That is, they should be clear, unidimensional, equally specific, achievable, and measurable.

For example, following are KAS components for the second objective above, "Students who are sexually active will have less unprotected intercourse by using effective forms of birth control."

Knowledge Areas

Students will know:

- the needs of children and the responsibilities and costs of parenthood
- basic facts of reproduction and fertilization
- the important characteristics of the major effective methods of birth control (e.g., the name, effectiveness, appropriate use, advantages and disadvantages, cost, source, and relevant laws)
- the reasons teenagers fail to obtain and use an effective form of birth control
- the consequences (physical, emotional, and social) of adolescent pregnancies.

Attitudes

Students will believe that:

- they are probably capable of becoming pregnant or impregnating if they are sexually active
- it is better to prevent an unwanted pregnancy than to have to deal with one
- both sexual partners have responsibility for preventing pregnancy and both should take that responsibility
- it is important to discuss the possibility of pregnancy and the use of birth control with a partner before becoming sexually active.

Skills

Students will be able to:

- refrain from having sex if they cannot use some effective form of birth control
- obtain an effective method of birth control
- use that method of birth control effectively.

Steps within Each Task

Following are eight steps that will help insure the comprehensiveness and selection of the important goals, objectives, and KAS components. You can use these steps with experts, members of your staff, community members, potential participants, and other appropriate people.

Step 1: Generate a comprehensive list of goals (objectives, KAS components).

Step 2: Continue giving this list to other experts until no additional goals (objectives, KAS components) are added.

Step 3: Organize the list in a logical manner (e.g., group items on a similar topic together).

Step 4: Give the list to a panel of experts and have them rate each item on a numerical scale (e.g., 1=not at all important, 2=slightly important, 3=somewhat important, 4=very important, 5=critical to the success of the program).

Step 5: Calculate the mean score for each item.

Step 6: Send the panel of experts the following information: which items received low ratings and should be excluded, which items received mixed ratings and should be discussed, and which items received consistently high ratings.

Step 7: Hold a meeting of the panel of experts; have them discuss each item; give them the opportunity to rewrite and reorganize items; and have them vote a second time on each item.

Step 8: Calculate the mean score for each item; include only those items that are clearly important.

If it is not possible to hold a meeting of the entire panel of experts, then hold a meeting of a smaller number of the experts. If this is not possible, then reorganize the items yourself, and send them with the ratings to the panel for a second vote.

In general, this entire process takes much longer than people typically estimate. For example, carefully specifying the important goals, objectives, and KAS components for a comprehensive program may require months or even a year. The less comprehensive the goals, and the more centrally located the panel of experts, the more quickly the process can be completed.

Although these steps are time consuming, they are worth the effort, because they will facilitate a much clearer and more precise specification of the goals, objectives, and KAS components. Once these things are specified, writing the curriculum and questions for questionnaires is relatively easy. For example, if the effectiveness of different forms of birth control has been specified as important, then you need to include that specific topic in the curriculum and you need to write a question that measures knowledge about contraceptive effectiveness.

Task 4: Identify Unexpected or Undesired Effects

The preceding discussion has focused upon specifying the important goals, objectives, and KAS components of programs. However, your program may sometimes have unexpected and negative effects, and these should also be measured. For example, critics of sexuality education have argued that sexuality education programs suggest new kinds of sexual activity to students, destroy students' morality by making them believe that premarital sex is acceptable, and increase students' sexual activity. Although people in this country vary greatly in their views of these possible outcomes, few programs have these outcomes as goals, and most programs would consider them unexpected or undesired outcomes.

Such unexpected or undesired outcomes should be measured for at least three reasons. First, your evaluation will be biased if you measure only possible desired consequences and ignore possible undesired consequences. Second, the program staff need to know if programs do in fact have undesired consequences, so that they can remedy the problems. Third, if programs do not have such an impact, this should also be documented so that there is evidence with which to respond to criticism and concerns.

CHAPTER 3

USING EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS

Chapter 1 compared the use of experimental designs and survey methods in the evaluation of sexuality education programs and argued that experimental designs provide much stronger evidence for the causal impact of programs. This chapter will describe several different experimental and quasi-experimental designs. It will discuss the most rudimentary of experimental designs, describe a major weakness of that design, provide a solution, describe another weakness, provide another solution, and continue until the design is adequate. This method of presentation is intended to demonstrate the rationale for each part of a good experimental design. The chapter will then describe other problems and other experimental designs that are useful in particular situations.

This chapter uses the symbols and terminology employed by Stanley and Campbell in their popular book Experimental and Quasi-experimental Designs for Research. You should read that book if you need a fuller discussion of experimental designs. In the diagrams:

- "X" represents the experimental treatment; in our case, it represents participation in a sexuality education course.
- "O" represents some observation, measurement, or testing; in our case, "O" represents the administration of questionnaires.
- "R" represents the random assignment of people to the experimental group (those who participate in the program) and to the control group (those who do not participate in the program).

Each row of symbols represents a different group of people. The left to right order of the R's, X's, and O's indicates the temporal order of the random assignment to the experimental or control groups, the participation in the program, and the observations (administrations of the questionnaire). Symbols in the same column indicate that those groups participate in the event (randomization, participation in the program, or observations) at the same time.

Throughout the following discussion of different experimental designs, this chapter will assume that the experimental treatment is participation in some type of sexuality education course and that the observations are some type of test. However, you should remember that the same principles apply to all types of experimental treatments and to all types of systematic observations (knowledge tests, other kinds of questionnaires, or other kinds of data such as pregnancy rates).

One-Shot Case Study

This is the most rudimentary of all experimental designs and perhaps it should be called a pre-experimental design. It contains the minimal components of an experimental design: participation of a group of people in an experimental treatment (a sexuality education class) followed by one observation or posttest.

Experimental Group X 0

Despite the rudimentary nature of the design, educators probably use this design more than any other design to evaluate their courses. Many educators teach their students some material and then test them at the end of that unit. If the students perform well on the tests, the teachers believe that the students have learned the material and that they have taught the material well. For example, if a group of students answer correctly 90 percent of the questions on a knowledge test, then the teacher may feel the course is effective.

Problem: Failure to measure change. This design has several flaws. Its critical flaw is that it fails to measure how much the students actually learned. For example, even if the students performed well on the posttest, the course may have been totally ineffective; the students may have known just as much before the course as after the course.

Solution: Administer tests before and after the sexuality education course in a One-Group Pretest Posttest Design.

One-Group Pretest Posttest Design

In this design students complete a pretest before the course, then take a course, and finally complete a posttest after the course.

Experimental Group 0 X 0

By measuring the difference between the pretest scores and the posttest scores, the researcher can measure the change that took place during the course. Consider the example below:

Experimental Group 70% X 90%

In this example, the students answered correctly 70% of the questions on a knowledge test before the course, 90% after the course. Thus, their test scores show a change or improvement of (90% - 70%) or 20%. This suggests that the course was effective.

Consider a different example:

Experimental Group 70% X 71%

In this example the students answered correctly 70% of the questions on the pretest and 71% of the questions on the posttest. Their improvement was only (71% -70%) or 1%, suggesting that the course did not effectively increase knowledge.

Problem: Failure to link change to the course. This quasi-experimental design also has a number of major flaws. The most critical flaw is that the measured change may not have been caused by the course but may have occurred anyway. For example, teenagers in high school are in a stage of rapid change. Regardless of whether or not they participate in a sexuality education class, they are likely to become more interested in sexuality, to learn more about sexuality, and to participate in various social and sexual behaviors for the first time.

Consequently, if researchers observe the change only in the students who take the class (the experimental group), they may incorrectly conclude that the increase in knowledge was caused by the course when in fact it would have occurred anyway. Similarly, they may incorrectly conclude that a course caused students to engage in sexual activity, when the students would have engaged in sexual activity anyway. Thus, the lack of a control group is obviously a major problem.

The seriousness of not having a control group depends partly upon the length of time between the pretests and the posttests and partly upon the occurrence of any special events during that elapsed time. For example, if a course is a semester or a year long, then students may learn a significant amount about sexuality and a few students may become more sexually active regardless of whether or not they participate in a sexuality education course. For such an extended time period, a control group is needed.

On the other hand, if a course is short and the pretests and posttests are only a couple of weeks apart or less, then a control group may or may not be needed. For example, if researchers are studying knowledge, if the pretests and posttests are only two weeks apart, and if the students did not participate in any special event, then the researchers can probably conclude 1) that the students would not have increased their knowledge score significantly if they had not participated in the course, and 2) that the course produced any significant improvement between the pretest and posttest scores.

If the researchers are studying attitudes or behavior, however, and if the students participated in parties after a major football game, attended a senior prom, or spent Easter vacation in some romantic spot, then these other special events and not the sexuality education course may have produced the changes in attitudes and behavior between the pretests and the posttests. In general, researchers should always use a control group if special events, normal maturation, or any factor other than participation in the sexuality education course may have affected the students' scores.

Solution: To overcome these problems, researchers should use a Nonequivalent Pretest Posttest Control Group Design.

Nonequivalent Pretest Posttest Control Group Design

In this design both the experimental and control groups complete the pretest. The experimental group then participates in a sexuality education course, and after the course, both the experimental and control groups complete the posttest.

Experimental Group	0	X	0
Control Group	0		0

The strength of this design is that it enables the researcher to compare the change in the experimental group with the change in the control group.

To illustrate this, consider the example below:

Experimental Group	70%	X	90%
Control Group	71%		72%

In this example, the experimental group increased its score by 20% while the control group increased its score by only 1%. This definitely indicates that the sexuality education class increased the knowledge of the students in the course.

Consider a second example:

Experimental Group	75%	X	91%
Control Group	72%		87%

In this example, the experimental group increased its percentage of correct answers by 16%, and the control group increased its scores by 15%. Because these increases are approximately the same, some factor other than the sexuality education class was probably responsible for the increase in knowledge, and the data indicate that the sexuality education class was not effective.

Finally, consider a third, somewhat common example:

Experimental Group	75%	X	91%
Control Group	60%		63%

Problem: Dissimilar control and experimental groups. Because this design does not include any procedure for assuring that the two groups are equivalent before the course, the two groups may differ. In the example, the experimental group increased its percentage of correct answers by 16% and the control group improved its scores by 3%. This comparison alone would indicate that the course was effective. However, the data indicate clearly that the experimental and control groups were not similar prior to the sexuality education class. In particular, the experimental group was already better informed than the control group. This indicates that the students who signed up for the sexuality education class were different from the other students. The program may have been effective for them, but may not have been effective for other students. Alternatively, the control group may have had some special quality that prevented them from learning a normal amount about sexuality from everyday life. If so, then they were not an adequate group with which to compare the experimental group.

The experimental and control groups may differ in other, unknown ways. In the example, the dissimilarity between the control and experimental groups is obvious; the pretests indicate that the groups have significantly different scores on the measured variables. However, even if the two groups have similar scores on the measured variables, they may nevertheless differ substantially on other unmeasured variables. Thus, even if the pretest scores are similar, the control group may not be an adequate control group.

Solution #1: One way to assure that experimental and control groups are similar is to try to find a control group that is as similar as possible to the experimental group. For example, one can use as a control group another class in the high school which has the same age distribution, the same grade level, the same level of capabilities and intelligence, and other similar characteristics.

Finding such a class may be difficult, because students who decide to take a sexuality education course may be different from those who don't. For example, they may be more liberal or more sexually active. If students going to college have a full schedule, they may have greater difficulty fitting sexuality education into their schedule, and thus sexuality courses may have fewer college-bound students.

Solution #2: Another approach is to match each student in the experimental group with a similar student for a control group. For example, if one person who signed up for the sexuality education class is very bright, male, Black, and 15 years old, then the researcher would try to find another very bright, male, Black, 15-year-old student who had not taken the course and add this person to the control group. Although the procedure greatly improves the similarity of the experimental and control groups, implementing it may be time consuming and difficult, and the groups may still differ on unmeasured variables.

Solution #3: The preferred method is to use a Randomized Pretest Posttest Control Group Design.

Randomized Pretest Posttest Control Group Design

This is the classical experimental design. It is the same as the former design except that the students are randomly assigned to the sexuality education class and the control group.

Experimental Group	R	O	X	O
Control Group	R	O		O

The advantage of this design is that randomly assigning students diminishes the differences between the experimental and control groups, and thus the control group is an excellent group with which to compare the experimental group.

Data is analyzed in the same manner as in the former design.

Problem: Impracticality. The major problem with this design does not involve the validity or strength of the conclusions, because this is an excellent design. The major problem is a practical administrative problem. Rarely can researchers randomly determine whether any individual must take a sexuality education class or must participate in a control class. For both moral and political reasons, researchers should neither force students to take a sexuality education class nor restrain them from taking the class.

Solution: Use the Delayed Treatment Design.

Delayed Treatment Design

In this design, students who sign up for a sexuality education course are randomly assigned to experimental and control groups.

	Phase 1				Phase 2	
Experimental Group	R	O	X	O		
Control Group	R	O		O	X	O?

By considering only those students who sign up for the course, we solve the dilemma of forcing students to take a sexuality education class. The students in the control group do not take the sexuality education course during Phase 1 of the study (to the left of the vertical line). During that phase they serve as a true control group by taking the pretest and posttest but not the course.

Then after the posttest and during Phase 2 (to the right of the vertical line), the control students can take the sexuality education course. This solves the dilemma of preventing them from taking the course. If desired, the researcher can administer a third questionnaire to the "control group" after they participate in the course. The third questionnaire can serve as a posttest to the second questionnaire which now serves as a pretest. That is, the control group can provide both control group data during Phase 1 and experimental group data during Phase 2.

Problem: Immediate versus longterm effects. All the designs discussed thus far measure the effects of the program at only one point in time, typically immediately after the program. This is a significant limitation because effects may change over time. Numerous studies demonstrate that students cram for final exams, take the exams, and then quickly forget a substantial part of the material. Thus, a posttest completed immediately after a course may exaggerate the long term impact of the course.

On the other hand, some effects may not occur until months or years have passed. For example, a sexuality education course that stresses the importance of getting prompt medical attention if a person has any signs of sexually transmitted disease (STD) may not have any behavioral consequences until months or years later when some of the students get STD. Similarly, a course that emphasizes using some effective form of birth control when sexually active cannot have any behavioral consequences until the students become sexually active. If a posttest questionnaire asking questions about these behaviors is administered immediately after a course, it might indicate that the course had no behavioral effects, when in fact, the course would affect subsequent behavior.

To overcome this limitation, a Pretest and Multiple Posttest Control Group Design should be used.

Pretest and Multiple Posttest Control Group Design

This design is the same as the Randomized Pretest and Posttest Control Group Design except that it includes a second posttest. This design provides the best and most valid data.

Experimental Group	R	0	X	0	0
Control Group	R	0		0	0

Two examples will illustrate the interpretation of the data.

Experimental Group	68%	X	90%	85%
Control Group	69%		70%	71%

In this example, the two groups of students began with approximately the same knowledge. The sexuality education course increased the experimental students' knowledge substantially, (90% - 68%) or 22%, and the control group remained about the same. Over a period of time, the students who took the course forgot some material. The increase dropped, (85% - 68%) to 17%, but they still knew substantially more than the control group.

Experimental Group	70%	X	90%	80%
Control Group	71%		76%	79%

The second set of data indicates that the sexuality education course temporarily increased experimental students' knowledge, but that the normal life experiences of the control group also gradually increased their knowledge; over a period of time, the two groups were again indistinguishable (80% compared to 79%). Thus, the sexuality education class had no long term effect.

Problem: Administrative difficulties. Most researchers have difficulty keeping contact with both the experimental and control groups over an extended period of time. Students move away, graduate from high school, drop out of school, or become lost from the sample for other reasons. Still others may remain in the sample but refuse or simply fail to return questionnaires after the end of the course. If a substantial percentage of respondents fail to complete one or more of the three questionnaires, then the mean scores may become very misleading and biased.

Solution #1: Match individuals' questionnaires. A partial solution to this problem is to remove from the analysis all questionnaires for anyone who did not complete all three questionnaires. For example, if a person completed the pretest but neither posttest, then the pretest questionnaire would be excluded from analysis. Alternatively, if a person completed the pretest and only the first posttest, those questionnaires could be included in the analysis of the short term effects, but they would be excluded in the analysis of the long term effects.

This solution requires some method of linking the pretest questionnaire for each person with the corresponding posttest of each person. An obvious method is to ask people to put their names on the questionnaires. If there is no need for anonymity, this is a good solution. However, in many analyses of sexuality education, sensitive questions are asked, and anonymity must be assured. Thus, including names is not allowed. An alternative strategy is to ask each student to put on the paper some meaningful number that

- is unique to that student
- can be remembered easily by that student
- is anonymous.

If all the students have social security numbers, then they can use the last four or five numbers in their social security number. Alternatively, they can use the four digits representing the month and day (but not the year) of their birthday. This method works well with small groups. If there are many students, however, more than one student may have the same birthday, but specifying the year in which they were born would reduce anonymity of either younger or older students. Yet another possible number is the last four or five digits of their phone numbers. In sum, identification numbers must be chosen with care.

This solution will prevent error caused by one group of people taking a pretest and different groups of people taking posttests. However, another problem may remain; those students who complete all the questionnaires may be significantly different from those students who fail to complete one or more of the questionnaires. If this occurs, your conclusions would apply to only those people who complete all the questionnaires and not to all the class participants.

Solution #2: Conduct a Follow-up of Nonrespondents. The best way to determine whether the course had a different impact upon those students who completed the evaluation and those who did not is to carefully track down some of the students who dropped out of the evaluation and to then compare them with the students who remained in the evaluation. If students dropped out because they found the course boring or because they became pregnant, then omitting them from the evaluation may significantly distort your results. On the other hand, if they dropped out because their parents moved but appear similar in other ways to the students in the evaluation, then omitting them from the evaluation may not bias your conclusions.

Problem: Testing Effects. When courses are short, taking a pretest may increase the amount students learn and may also improve their ability to complete the posttest. For example, after students complete a knowledge test and listen to a lecture on the material covered in the test, they may remember the questions on the test and pay special attention to material that answers those questions. That is, they may selectively learn the material that was covered on the pretest. If this happens, and if the posttest is the same as the pretest, then the difference in scores between the pretest and the posttest may overstate the amount learned by the students. Similarly, if after completing a test, students ask other students about correct answers to some of the questions, then once again, the change in test scores may overstate the benefits of the course. Finally, if the pretests and posttests are administered only a short time apart and if they are the same tests, then on the posttest students may simply remember their correct answers on the pretest and may be able to devote more time to the questions that were more difficult for them.

First Solution: Use the Posttest-Only Control Group Design.

Posttest-Only Control Group Design

In this design the students are randomly assigned to the experimental and control groups; the experimental groups participates in the course; and then both groups complete questionnaires after the course.

Experimental Group	R	X	0
Control Group	R		0

This design eliminates the effects of pretesting by eliminating the pretests. This also reduces the total time required for test administration. If the randomization is completed carefully and if the sample sizes are large (e.g., more than 100), then the experimental and control groups should be very similar to each other before the course begins, and any observed differences after the course should be caused by the course. Moreover, additional posttests can be administered to measure long term effects.

The major disadvantage of this design is that the sample size must be reasonably large and the randomization must be completed properly. If either of these conditions is not met, then the two groups may differ significantly prior to the course; this difference would not be measured; and it might be incorrectly attributed to the course.

Second Solution: Use different but equally difficult tests, using the Alternative Tests Design.

Alternative Tests Design

In this design students are randomly divided into two different experimental groups and two different control groups. One experimental group and one control group get one questionnaire, 0′, as a pretest and a second questionnaire, 0′′, as a posttest. The remaining two groups get the two questionnaires in reverse order.

Experimental Group	R	0′	X	0′′
Experimental Group	R	0′′	X	0′
Control Group	R	0′		0′′
Control Group	R	0′′		0′

If possible, the two versions of the questionnaire should be equally difficult. However, even if the two versions are not equally difficult, this design still works if students do not drop out and the same number of students take both versions as a pretest and the same number take both versions as a posttest. Then the scores for both experimental groups can be combined (or averaged). Similarly, the scores of the two control groups can be combined (or averaged).

Third Solution: Use the Solomon Four Group Design.

Solomon Four Group Design

In this design one of the experimental groups and one of the control groups complete both the pretests and the posttests, while the other experimental and control groups complete only the posttests.

Experimental Group	R	O	X	O
Experimental Group	R		X	O
Control Group	R	O		O
Control Group	R			O

This design enables the researcher to measure directly the impact of the pretest contamination. The design is based upon the following principle: if students are carefully randomly assigned to groups, then the mean scores on the pretests would be the same for all four groups if all four groups had completed the questionnaires. If the posttest scores of the experimental (or control) group with the pretest differ from the posttest scores of the experimental (or control) group without the pretest, then the pretest may have affected the posttest scores.

Consider the following example:

Experimental Group	70%	X	90%
Experimental Group		X	85%
Control Group	70%		74%
Control Group			70%

This data suggests that the pretest increased the scores in the experimental group by about 5% and increased the scores in the control group by about 4%.

Several researchers have used the Solomon Four Group Design to measure the impact of pretesting. Their preliminary conclusion is that pretesting has little impact if questionnaires are administered at least a week apart. If they are administered in the same day, pretesting may or may not have an impact.

Problem: Control group contamination. In some schools or youth agencies, the students in the experimental group may interact with the students in the control group. If so, one group may "contaminate" the other group. For example, if a sexuality education class has emphasized the importance of using some form of birth control, and a member of the sexuality education class is sexually involved with a member of the control group, the impact of the sexuality education class may affect both people. Alternatively, the lack of sexuality education for the control group member may affect both of them. In either case, error is introduced into the data. This error is particularly significant if the school or youth group is small, and friends discuss the contents of the course with one another.

Solution: There is no design that overcomes this problem. Instead, the researchers need to select control groups that have little interaction with the experimental groups. If control group contamination appears to be a substantial problem, then the researchers should consider obtaining a control group from some other school or population of young people. This, of course, introduces new problems, because random assignment is then difficult and the control group in a different school or population may not be similar to the experimental group.

Problem: Inability to obtain a control group. In many situations, obtaining an adequate control group is difficult, if not impossible. For example, everyone in a school may take a sexuality education program at the same time and consequently, no one is left to serve as a control group. Or more realistically, many people in a school may take a sexuality education program and those who fail to take it are not similar to those taking the course.

Solution: Improve the Single Group Pretest Posttest Design by increasing the number of different pretests and posttests.

Time Series Design

This design lacks a control group, but it does include several pretests and posttests.

Experimental Group 0 0 0 0 X 0 0 0 0

In general, the larger the number of pretests and posttests, the more valid the conclusion. The additional pretests and posttests allow the researcher to establish a more solid basis before and after the sexuality education course, and therefore, to make a more conclusive claim about the effects of the course. Ideally, the time that elapses between the last pretest and the first posttest is similar to the time between the other pretests or posttests.

Consider the following example:

Experimental Group 70% 72% 71% 72% X 85% 84% 86% 84%

Assuming equal time periods, the stability of the scores before the sexuality education course, the sudden increase in the scores during the course, and the stability of the scores after the course strongly suggest that the course and not normal maturation processes produced the change. Of course, the possibility that some other major event affected the scores still remains.

Consider a second example:

Experimental Group 65% 70% 74% X 90% 95%

In this example, the effects are less clear. The scores increase both before and after the course, but they increase more during the course. This indicates that the course did have an effect.

Finally, consider a third example:

Experimental Group 65% 75% 66% X 77% 67% 80% 65%

In this example, the scores vary so much from one test to the next that one cannot conclude that the course produced an increase even though the scores increased 11% between the last pretest and the first posttest.

If the Time Series Design is used with questionnaires, students may stop answering them carefully after the first few administrations. Moreover, the effects of testing may become substantial.

The Time Series Design is most often used when pregnancy rates or other data are collected over time by some outside agency. For example, a school may implement a sexuality education program and afterwards obtain estimates from nearby clinics of the number of pregnancies in that school for each of several years before and after the program was implemented.

Although this approach is sound in principle, it has three problems in practice. First, the pregnancy rates for schools (or other groups of people) often vary considerably from one year to the next. Thus, it is difficult to distinguish between changes produced by the program and normal variations in pregnancy rates. Second, programs are often implemented gradually over time. During the first couple of years, if only 10% of the school participates in the program, the pregnancy rate would be decreased by only 10% even if the program were perfectly successful in preventing pregnancies, and this small amount of decrease could be obscured by the normal amount of change from year to year. Third, the time lag of a program must be estimated. Even if everyone participates in a program, the effects may not be immediate. Thus, once again, it will be more difficult to separate changes caused by the program from changes caused by other factors in the community.

Summary

The great advantage of experimental designs is that they increase the ability of the researchers to control:

- the assignment of students to experimental and control groups
- the design and content of the sexuality education course
- the relative timing of the testing and the course.

This, in turn, greatly increases the ability of the researchers to compare:

- pretests and posttests
- multiple posttests
- experimental groups and control groups.

Finally, this ability greatly increases the validity of statements about the causal impact of programs. Few other designs can provide such solid evidence about causality.

You should select the best design that is feasible in your circumstances. At a minimum, you must administer pretests and posttests to an experimental group. If at all possible, you should have a control group. If you are measuring behavior, you should definitely try to measure long term effects with additional posttests. If you have a control group that is likely to be very similar to your experimental group, you probably do not need to worry as much about randomly assigning people to the experimental and control groups, although you should do so if possible. Similarly, you probably do not need to worry about pretest contamination and thus do not need to use a Solomon Four Group Design.

CHAPTER 4

CONDUCTING SURVEYS

When conducting a survey, researchers typically collect information from a sample of people. To do this, they usually:

- specify the problem
- select the basic paramaters of the survey
- identify the important variables to be measured
- design interview schedules, questionnaires, or other methods of collecting data
- select a sample
- conduct personal or telephone interviews, administer questionnaires or use other methods of collecting data
- analyze the data statistically.

This chapter will briefly discuss 1) different kinds of surveys, 2) their advantages and disadvantages over experimental designs, and 3) ways to use survey methods to evaluate sexuality education programs. Other chapters cover other important topics in survey research: identifying important variables, designing questionnaires, selecting a sample, administering questionnaires, and analyzing data.

Although large surveys can be used to evaluate sexuality education, this handbook is definitely not designed to prepare you to conduct a large survey. If you intend to conduct such a survey, you should read a textbook on survey research.

Types of Surveys

Methods of Collecting Survey Data

Personal interviews. Some of the best examples of survey research have used interviews based upon detailed interview schedules or questionnaires. During the interviews, the interviewer can make sure that all questions are understood and answered and can ask additional, more detailed or probing questions when necessary. When interviewers are properly trained, they are especially likely to obtain complete and valid data. However, interviews are less appropriate in evaluating sexuality education programs; they are not anonymous and young people may be unwilling to be honest when answering sensitive questions about sexuality. Moreover, interviews are time consuming and costly.

Telephone interviews. It is normally much quicker, easier, and cheaper to interview people on the telephone than in person. However, telephone interviews do have several limitations: respondents are less likely to cooperate or to answer personal or sensitive questions, and the interviews must be short. Telephone interviews are sometimes used to contact a random sample of the students′ parents to ask them their views of the course and its effects upon the students.

Questionnaires. Written questionnaires are probably the best method of collecting survey data that includes any potentially sensitive information. They are very commonly used in both large and small surveys.

Survey Designs

One-shot surveys. In most surveys, information is collected at only one point in time. Thus, if you intend to use a survey to measure the impact of a program or reaction to a program, you should complete the survey after the program ends. To measure the impact, you should collect information from both participants and nonparticipants.

Panel surveys. In panel surveys, information is collected from the same group of people at more than one point in time. This enables the researcher to measure change in the people over time.

Surveys Versus Experimental Designs

Random Assignments

In general, the distinguishing characteristic between an experimental design and a survey is that an experiment includes the random assignment of people to experimental and control groups and a survey does not. However, there are many kinds of experimental and quasi-experimental designs and many kinds of surveys and they can overlap. For example, one of the quasi-experimental designs discussed in the previous chapter did not include random assignment to experimental and control groups and did include giving questionnaires to experimental and control groups before and after the program. This is identical to a panel survey in which questionnaires are also given to experimental and control subjects before and after a program. In other words, as you take away some of the characteristics of true experiments and add to the characteristics of surveys, the two converge and what you label them makes relatively little difference.

True experimental designs with random assignment can provide the most compelling evidence about the causal impact of sexuality education programs. This is a great advantage of experimental designs, but also a major burden. For example, to evaluate a specific program with an experimental design, you may need to obtain from the school authorities and the teachers involved their approval of the evaluation, the experimental design, and the random assignment. You also need the consent and cooperation of the students to be randomly assigned to the two groups and to complete one or more questionnaires. Finally, you may need permission from the parents to allow their young people to participate in the program and the experimental design.

In contrast, to administer a survey, you simply need the cooperation of the respondents to complete one questionnaire and you may need the permission of the parents for their students to complete the questionnaires. Obviously, completing the survey is much easier and requires less control than the experimental design.

Sampling Flexibility

A second major advantage of surveys is their much greater sampling flexibility. When using an experimental design, you must administer questionnaires to the group of people who participated in a program and preferably to an additional similar control group as well. When conducting a survey, you can administer questionnaires to many different groups of people, although if your goal is to evaluate the impact of sexuality education programs, you must include in your sample some people who have taken sexuality education and some who have not.

The flexibility inherent in surveys is demonstrated in previous research. A few studies have administered questionnaires to all students in a high school and compared those who had taken sexuality education with those who had not. Other studies have selected random samples of all teenagers in the United States and then compared those who had taken sexuality education with those who had not. Still others have taken surveys of program participants, parents of program participants, and members of communities with new sexuality education programs.

Demonstration of Causality

The major disadvantage of surveys is that they cannot fully control other relevant variables and thus cannot provide compelling evidence for the causal impact of sexuality education programs. For example, a survey of high school students might reveal that those students who have taken sexuality education have also been more sexually active. However, this conclusion would not necessarily result from sexuality education courses causing greater sexual activity. Rather, it could result from two other relationships: as high school students get older they are 1) more likely to take sexuality education and 2) more likely to have become sexually active. Thus, sexuality education would not have caused the sexual activity; rather, it would have been caused by the third factor, age.

Although it may be possible to control for age statistically, other factors may produce spurious relationships that could not be controlled for. For example, students who choose to take sexuality education courses may be more predisposed to become sexually active or may be more sexually active even before taking the course than students who do not choose to take the course. Unlike experimental designs, surveys cannot control for such factors.

Surveys that are being used to measure the impact of sexuality education programs in general have another major problem. Respondents may have participated in numerous different sexuality education programs, about which the researchers may have little information. Including some questions in the questionnaire about the quality and comprehensiveness of the programs would still provide insufficient information about the programs.

Uses in Designing and Evaluating Programs

Although surveys have serious problems, they do have several important uses.

Assessing Needs

When educators design sexuality education courses, they frequently find it useful to determine what the potential participants need by conducting needs assessments. This entails sending a sample of potential participants a questionnaire with the following kinds of questions:

- questions about the background of the participants
- lists of topics for participants to rate according to their importance to the participants
- open ended questions requesting additional topics that should be covered.

Sometimes educators demonstrate need by designing a knowledge test with questions about facts that young people need to know, administering it to young people, and then demonstrating that most of the respondents missed many questions.

Assessing Community Opinion

Unfortunately, sexuality education programs are often controversial, and a small number of vocal people may gave a misleading impression that there is a lot of support or opposition to a program. To accurately determine the actual amount of community support or opposition to programs, some educators have surveyed the community.

When implementing such a survey, you should be sure to send questionnaires to either all parents (or members of the community) or to a random sample of the parents (or members of the community). You should also try to get a high response rate so that the survey data accurately reflect community opinion. If only those people who strongly favor or strongly oppose sexuality education complete and return the questionnaire, then the data will be biased. Response rates of 80 or 90 percent are excellent, but unusual. Researchers can normally obtain rates of 50 or 60 percent at most. If your response rates are lower than 80 or 90 percent, you should try to determine how the respondents differed from the nonrespondents.

When assessing community opinion, you may find it useful to provide a list of topics that might be covered in a course and then ask parents (or community members) to indicate which topics should be covered and in which grades (5th grade, 6th grade, etc.) they should be covered. If you are sending questionnaires to members of the community in general, you may want to include a question asking whether they have children in the school.

Assessing Immediate Reaction to Programs

As discussed in the first chapter of this volume, formative evaluation provides immediate feedback to the instructors about participants' reaction to the instruction. Such feedback is often very important, because it allows instructors to immediately change or modify the course even before it has been completed. It is also an important addition to verbal feedback, because it is anonymous.

An important way to obtain feedback from students is to administer questionnaires at frequent intervals throughout the course. These questionnaires can contain the following kinds of questions:

- lists of different topics with instructions to rate the value of each topic
- open ended questions asking students to suggest additional topics
- open ended questions asking students to suggest any kind of change or improvement

- characteristics of the teacher with instructions to rate the teacher on each dimension
- characteristics of the classroom environment with instructions to rate the classroom on each dimension
- questions about how the students feel the course has or will affect them.

Questionnaires can be administered in class, distributed in class as homework, distributed through other organizations, or mailed. However, as many teachers know, if the questionnaires are not completed and collected in class, many people will not return them.

Examples are found in the appendix.

Assessing Parent and Community Reaction to Programs

Because parents have the opportunity to observe their children's behavior off campus, they can provide an independent view of the impact of the program. One approach is to send questionnaires to parents at the end of the program, asking them about their reactions to the program. Such questionnaires should ask for the following kinds of information:

- amount of parental involvement in the course
- amount of parental knowledge about the course
- parents' perception of the effects of the course upon their children
- degree to which the parents support the program
- parents' suggestions for any changes in the program (open ended questions).

Once again there are examples in the appendix.

Assessing the Effects of the Program

There are two ways to use surveys to assess the effects of programs. The first is to administer questionnaires to people who have taken one or more programs and also to people who have taken no sexuality education programs. This approach has the disadvantages discussed above, especially the researcher's inability to control relevant variables, and the lack of detailed information about the programs taken.

The second way is to administer questionnaires to students themselves and ask them for their perception, namely, how they believe the course has or will affect them. This method can provide some useful supplementary information. However, by itself, it is not a valid method for several reasons. First, if students like the program, they will claim that it had very positive effects even if it did not. Second, the method relies upon the students' subjective perceptions, which may be incorrect. They may believe the course will change their behavior, when in fact their behavior will not change or would change anyway.

Conclusion

In sum, surveys can elicit useful information from different people during the planning and revision of programs. They cannot provide the compelling data that can be provided by experimental designs for the causal impact of programs, but they can provide very useful supplementary information.

CHAPTER 5

EMPLOYING THE FUNDAMENTALS OF QUESTIONNAIRE DESIGN

This handbook devotes considerable space to the design and use of questionnaires to evaluate sexuality education programs. There are three reasons for this. First, most evaluations of sexuality education programs have used questionnaires. Second, anonymous questionnaires usually provide the best method of collecting new data on the numerous possible outcomes of sexuality education programs. Third, other methods may not be feasible. For example, although a program may attempt to influence students' social and sexual behaviors, you cannot directly measure those behaviors outside of the classroom. Similarly, interviewing is not only very time consuming, but sacrifices the respondents' anonymity and probably the validity of the data as well. In sum, anonymous questionnaires are usually the most practical way to obtain reliable and valid data on the effects of sexuality education.

Different designs require different kinds of questions. For example, if you want to measure the impact of the program upon the participants and are giving questionnaires to the students only at the end of the program, the questions must ask the participants how they believe the program has and/or will affect them. In contrast, if you are giving questionnaires to the participants both before and after the program, the questions should measure their knowledge, attitudes, or behaviors so that you can compare their pretest scores with their posttest scores. Thus, different designs require different kinds of questions.

Chapter 6 discusses questionnaires that measure participants' assessments of the program, and that are best administered only at the end of the program. Chapters 7, 8, and 9 discuss questionnaires that measure knowledge, attitudes, and behavior, and that can best be administered both before and after the program.

Important Steps in Designing Questionnaires

To develop reliable, valid, sensitive, and appropriate questionnaires, you should complete the following steps:

1. Determine the features and outcomes to be measured.
   - Consider state guidelines, community standards, and sensitivities of the students.
   - Specify the number of questions needed to measure each feature and outcome.
2. Construct the questionnaire.
   - Review other questionnaires and select questions.
   - Select the best formats for the questionnaire (multiple choice questions, Likert scales, etc.).
   - Write or rewrite the questions.

- Organize the questionnaire.
- Review the questionnaire several days later.
- Review the questionnaire for sensitivity.

3. Pretest the questionnaire.

- Have other experts review the questionnaire.
- Pretest the questionnaire with a few students.
- Review the the overall distributions of answers.
- Pretest the questionnaire with a larger group of students.
- Analyze the responses to each question, and modity the questions if necessary.

4. Assess reliability and validity.

- Assess the reliability and validity of each question, and remove or modify those questions with poor reliability and validity.
- Make sure the final distribution of questions and answers is acceptable.
- Assess the reliability of the entire questionnaire.
- Assess the validity of the entire questionnaire.

Determining the Important Features and Outcomes That Should Be Measured

Chapter 2 strongly encourages you to develop the curriculum specifications and the test specifications simultaneously and to develop both of them before implementing the program. If you follow that plan, your program and your evaluation will probably be consistent.

Unfortunately, many program providers follow a different developmental sequence; they gradually develop and modify a program, then decide to evaluate it, and then specify its goals and objectives in order to conduct the evaluation. In doing this, they risk specifying new and unrealistic or unfair goals, and basing the evaluation upon these goals rather than upon the original goals of the program. Then they lack congruence between the actual objectives of the program and the objectives used in the evaluation. If, for example, knowledge tests measure facts or content areas that are not emphasized in the classroom, the questionnaires may underestimate the amount that the students learned. It is essential to ensure that program and evaluation objectives are consistent.

Considering Laws, Regulations, and Community Values

When you are deciding which features and outcomes of the program you wish to measure and evaluate, you need to consider any regulations pertaining to the administration of questionnaires to students. Although all states allow teachers to administer knowledge tests about sexuality, several states and many school districts have regulations governing the administration of questionnaires about sexual attitudes and behavior. Only a few districts forbid such questionnaires, but many place restrictions; they may require that parents have the opportunity to see the questionnaires, that parental permission be obtained, or that school boards approve the questionnaires. You should be certain to meet these regulations. If you fail to do so, your evaluation and your program may be endangered.

You should also consider the sensitivities of the students and the values of the community. There may be some outcomes you should not try to measure because either your students or your community would find the questions offensive or overly personal.

Specifying the Number of Questions to Ask in Each Area

After making final decisions about what features and outcomes you do wish to evaluate and measure, you often need to decide approximately how many questions you will need in order to measure each feature or outcome. For example, if you have specified several different knowledge topics that should be covered, you then need to decide how many questions you should ask in each area. Similarly, if you wish to measure self esteem, you need to estimate the number of questions that you will need in a self esteem scale. Approximate numbers of questions needed for different kinds of scales are discussed in Chapters 6-9.

For the first draft, develop a larger number of questions than you ultimately wish to ask in each area. During the pretesting some of the questions will probably be unreliable, invalid, or unusable for other reasons and will be deleted.

Sometimes you may want to measure many program outcomes and may need to ask more questions than you can reasonably expect students to answer. If you have a sufficiently large sample of participants, then you can use matrix sampling, a technique that allows you to ask a greater number of questions and thereby measure more of the salient outcomes without making any single questionnaire too long and overburdening the participants. To use this technique, divide all the questions you wish to ask into several different questionnaires. Then randomly divide your respondents into the number of groups that you have questionnaires. If the groups are completing both a pretest and a posttest, be sure that they complete the same questionnaire on both.

Constructing the Questionnaire

Reviewing Other Questionnaires

Before designing your own questions and questionnaire, you should review existing questionnaires. They can provide a useful pool of questions, some of which may be appropriate for your evaluation and may also be reliable and valid.

- Some questions will stimulate ideas about additional questions;
- Some can be modified to meet your own particular needs;
- Some can be used as they exist.

However, when reviewing questions and questionnaires created by others, you should remain cautious for several reasons. First, those questionnaires may not have been properly designed. Second, questions which are valid for another population may not be valid for your population because your program participants may differ in age, in educational level, or in some other fundamental way. Or your program may differ; it may have emphasized different facts, attitudes, or behaviors. It is easy to use questions created by others, but not always appropriate.

A few questionnaires that have been carefully developed for the evaluation of sexuality education programs are listed at the end of the appropriate chapters. In the Appendix are questionnaires that we have used in our evaluation of sexuality education programs.

Writing and Organizing the Questions

Writing good questions is an art at which most people improve with practice. However, the following guidelines can help you avoid common errors.

- Use a vocabulary that all the respondents can understand.
- Use simple sentence structure.
- Make questions as clear as possible.
- Avoid double negatives.
- Avoid any trick or misleading questions.
- Make questions (especially attitude and behavior questions), unidimensional; in any question, focus upon only one topic, attitude, or behavior.

  Poor: How often do you talk about sexuality with your friends and parents?

  Better: How often do you talk about sexuality with your friends?
  How often do you talk about sexuality with your parents?

- Make sure questions are appropriately open-ended or closed-ended. Closed ended questions provide all possible answers and ask the respondent to select the best of the answers (e.g, "What is your sex? ___Male ___Female"). Open-ended questions provide a blank space and ask the respondent to write the answer (e.g., "Describe something you liked about this course.")

Organizing an entire questionnaire is also an art guided by these basic principles:

- Provide clear, complete, and concise directions.
- Arrange questions so that they are easy to read (e.g., put each choice of a multiple choice question on a different line).
- Organize questions by format (e.g., all multiple choice questions together).
- Put easier questions first and more difficult questions last.
- Put closed-ended questions first and open-ended questions last.

Keep the entire questionnaire short. If at all possible, a person should be able to complete it within 20 minutes. Questionnaires for elementary school students should take less time; questionnaires for college students can take more time. If necessary, administer different portions of the questionnaire on different days or use matrix sampling.

Analyze the responses to be sure correct responses form a random pattern. Avoid several consecutive questions with the same answer. For example, in a knowledge test you should have no more than three consecutive multiple choice questions with the correct answer "a." Similarly, in attitude scales you should avoid strings of items to which the respondent is likely to "Strongly Agree" or "Strongly Disagree." If you discover such response sets, you should change the order of questions or change the correct answers to questions.

Finally, avoid having a disproportionate number of questions with the same answer. If you have a multiple choice test with five possible answers for each question, then about 20% of the questions should have "a" as the correct answer, about 20% should have "b", etc. Having as many as 40% with the same correct answer may affect the respondents' answers and may adversely affect the validity of the questionnaire.

Reviewing your draft. After you have prepared a draft of the questionnaire, set it aside for several days, and then review it. Normally you will find that your fresh look will give you new insights and enable you to make numerous improvements.

At this time, you should also review the questionnaire with the sensitivity and values of the students and community in mind. Modify or remove any questions that might:

- offend the students
- offend members of the community
- inadvertently teach students incorrect information or promote improper values and behavior.

In addition, if you intend to ask questions about personal behaviors that, if revealed, could harm the students, you should either not ask the questions or take all reasonable precautions to ensure that the information will remain anonymous.

Pretesting the Questionnaire

Evaluators have commented that the three most important tasks in writing a questionnaire are pretesting, pretesting, and pretesting. Certainly pretesting is a very important task that is not always fully completed. Our experience clearly demonstrates that pretesting questionnaires uncovers confusion or other problems that we could not have anticipated and thus greatly improves the final versions of the questionnaires.

Using Experts

When writing questions, have other sexuality educators or knowledgeable professionals complete the questionnaire, examine it for clarity, suggest improvements, and if it is a knowledge test, verify the correct answers.

Using a Small Group of Program Participants

In addition, administer the questionnaire to about 5-10 participants with different skills and backgrounds.

- Observe any problems that they have while taking the questionnaire.
- Measure the time they need to complete it.
- Discuss each question on the questionnaire one at a time with them.

During these discussions, you should ask them how they interpreted each question and whether the questions contained needlessly difficult words, were clear, or were too personal or embarrassing. Ask if there were any other sources of confusion and how the questions could be improved. Such a process, if done properly, can elicit numerous suggestions for improving the questionnaires.

Pretesting the Questionnaire with a Larger Group of Students

At this point, it is important to administer the questionnaire to at least 30 students. During the administration, you should observe any problems that the students have with the questionnaire. Then score the answers and conduct an item analysis.

Item analysis involves the examination of the responses to different questions. Because it is particularly important in constructing knowledge tests and attitude scales, this handbook discusses item analysis in those chapters. However, an item analysis is often useful with other types of questions to determine whether questions were too difficult or were misunderstood, and whether questions should be made more or less extreme. Be especially careful when using criterion reference methods. Removing items or making them more difficult or easy may destroy your criteria. For example, if you removed several items because they were difficult and most people missed them, you might erroneously conclude that students who subsequently took the test had a sufficient grasp of the material.

If you are creating scales, you may wish to use factor analysis or other statistical methods to select or improve the questions in the scales. This requires a sample size of at least 50 and is fully discussed in Chapter 8.

Assessing Reliability

Reliability is the consistency, replicability, or reproducibility of a question, scale, or entire questionnaire. A reliability test measures how well you are measuring whatever you are measuring; it does not measure how well you are measuring what you think you are measuring.

To realize the importance of reliability, consider as an example an ordinary bathroom scale that measures your weight. If you stood on it five times in quick succession, and each time it gave you very different weights, you would consider the scale unreliable. On the other hand, if it gave the same weight each time, you would probably consider it reliable, although not necessarily accurate or valid. Similarly, if you gave a group of students who were not in a related course a knowledge test several times during a short period of time and they received widely different scores each time, it would not be reliable. If the students gave similar answers each time, it would be reliable. More generally, if a questionnaire fails to provide reproducible results, then it is not reliable and probably not useful. If it does produce reproducible and consistent results, then it is reliable and may be useful.

The reliability of a questionnaire can be measured in several different ways. Some methods are better for particular kinds of questionnaires than others.

Test-Retest Reliability

In the test-retest method, as suggested by the name, the researcher gives the same people the same questionnaire twice and compares the results of each administration. Two principles guide the time interval between the two administrations. First, if respondents can remember their answers from the first administration and simply repeat them during the second, the second administration would not be an independent measure of the phenomenon being measured. Therefore, the two administrations must be sufficiently far apart. Second, if the phenomenon being measured changes between the two administrations, the results would change even if the test was reliable. Thus, the two administrations should be sufficiently close together that the phenomenon being measured has not actually changed. For many questionnaires, a duration of 2 weeks between the two administrations is a reasonable solution to these two conflicting principles. However, if the questionnaire is a knowledge test, and if the students are covering in class the material in the test, then 2 weeks is obviously too long. You should either shorten the time interval, or give both the test and retest to a different group of students not taking the course.

If you have a calculator with correlation or a small computer, an easy way to compare the scores from the first administration with those of the second is to calculate the correlation coefficient between the two administrations. If students who scored poorly on the first administration also scored poorly on the second, and students who scored well on the first also scored well on the second, the correlation coefficient will be high, and the test is reliable. Correlations in the high .80′s and .90′s represent high reliability; correlations in the .70′s and low .80′s represent adequate or fair reliability; and correlations below .70 reflect poor reliability.

If you do not have a calculator with correlation or a small computer, then you should visually compare each individual′s pretest and retest scores. If the scores appear similar, then the questionnaire is probably reliable. Of course, inspecting visually is much less precise than calculating a correlation coefficient.

Split-Half Method

In the split-half method, the researcher measures the same underlying phenomenon twice, but this time measures it during the same administration with slightly different questions. Thus, the questionnaire would contain two parts, each of which would measure all the content areas. In a knowledge test, for example, each content area could have two similar questions, one in each half. These pairs of questions should be so designed that if particular students knew the answer to one, they would probably also know the answer to the other. Then, instead of comparing the two administrations of the same test, you would compare the two different halves of the test. Again, an easy way to make this comparison is to calculate the correlation coefficient between the two halves.

Multiple-Item Method

The multiple-item method is actually an extension of the principles involved in the split-half method. Instead of containing only two similar questions, the questionnaire would contain several questions measuring the same thing. For example, when measuring attitude toward premarital intercourse, a researcher could create a five-item scale with each item or question measuring the student′s attitude, then compare all five responses from each student. Although it is not possible to calculate a single correlation coefficient among three or more questions, an excellent statistic called Cronbach′s alpha does summarize the extent to which all the questions presumed to measure the same phenomenon are interrelated. Cronbach′s alpha can be interpreted in the same way as a correlation coefficient to establish reliability and can be found in standard statistical packages such as the Statistical Package for the Social Sciences.

Reliability and Criterion Referenced Measures

Some criterion referenced measures may have little variation in the answers.

That is, many people may answer the question identically. When this occurs, test-retest correlations or multi-item reliability coefficients will be low even when the items are reliable. Thus, contrary to the discussion above, when the variation is low, you should not conclude that the item is not reliable, even if the correlations are low. On the other hand, if the correlations are high, you can still conclude that the item is reliable.

Assessing Validity

Validity measures how well you are measuring what you want to be measuring. This is in contrast to reliability which measures how well you are consistently measuring whatever you are measuring. Return for a moment to the example of the bathroom scale. If the bathroom scale is actually a thermometer in disguise and is measuring the room temperature, then it may be reliable because it consistently measures the same temperature, but it is not a valid measure of weight because it is not measuring weight. More generally, questionnaires may be reliable but not valid.

By definition, you want your questionnaires to measure the phenomena that you designed them to measure. Thus, the validity of questionnaires is very important. When collecting evidence to demonstrate that you are in fact measuring what you want to measure, you may find one or more kinds of validity.

Face Validity

Face validity, the simplest and most direct kind of validity, measures the extent to which a question obviously or "on the face" measures the desired concept. The examples below have high face validity, because people will probably interpret the questions as they are intended and will answer them honestly.

What is your sex? ____ Female ____ Male

In what month were you born? ___

How often do you talk with your parents about methods of birth control?

In contrast, the next examples have lower face validity, because they may be misunderstood, may be difficult to answer, or may be answered dishonestly.

How many times each year do you have sexual intercourse? ___

How effective was the program? ___ Very effective
___ Somewhat effective
___ Not at all effective

Some young people may not know the meaning of sexual intercourse, may not remember how many times they had intercourse, or may not be willing to answer the question honestly if they think others may see their answers. Hence, their answers may be invalid. The second question may be invalid because people who like their instructors tend to overrate the effectiveness of programs.

If you are including questions of a slightly more technical nature (e.g., self esteem items), you can ask an independent professional such as a psychologist to assess whether the items measure the desired concept.

Because face validity can be overused and is not scientific, many texts reject it as a serious type of validity. You should claim your questions have face validity only if you are certain of their face validity and have no alternatives.

Content Validity

Content validity involves the extent to which questions proportionately cover a specified domain. To demonstrate content validity, you must demonstrate

- that experts in the field have selected certain facts, attitudes, or behaviors that should be measured, and
- that you are actually including questions on those particular facts, attitudes, or behaviors.

It may be the most common type of validity that you will use, especially if you are using criterion referenced methods. If you have followed the procedures for developing criterion referenced methods and specified the important goals, behavioral objectives, and necessary facts, attitudes, and skills, you should be able to demonstrate a reasonably high content validity.

Content validity is most appropriate for knowledge tests and less appropriate for questionnaires where different respondents may interpret items differently, misunderstand items, or refuse to answer them.

Criterion-Related Validity

Criterion-related validity measures the extent to which you are measuring a desired construct by independently collecting valid data on the respondents and then comparing that data with their responses on the questionnaires. Because this method involves verifying questions against known and valid data, it is probably the best type of evidence for validity. (Note that criterion related validity is different from criterion referenced methods. Although they have a common theoretical underpinning, they are distinct and should not be confused.)

Criterion-related validity is difficult to implement in sexuality research. Sometimes you can compare results from your questionnaire with results from another questionnaire that is known to be valid. However, other than knowledge tests, few questionnaires on sexuality have been established as valid for young people. Or you might administer a questionnaire about use of birth control methods to young people in a clinic, and then compare answers on the questionnaire to their clinic records. This comparison could serve to validate the questions, and then you could use the questionnaire independent with another similar population. However, such opportunities are relatively rare.

Predictive Validity

The difference between predictive and criterion-related validity is simply a matter of timing. In criterion-related validity you compare your questionnaire results with data collected previously or simultaneously from a different and valid source. In predictive validity you use your questionnaire data to predict subsequent behavior and then compare your predictions with that behavior.

Because of this similarity, predictive validity has the same advantages and disadvantages as criterion-related validity; it is a compelling method but very difficult to implement.

Construct Validity

Construct validity also resembles criterion-related validity to a considerable extent. In criterion-related validity, both your questionnaire and the independent source must measure the same concept. In construct validity, they measure two different concepts that are theoretically related. That is, you hypothesize that your questionnaire data will have certain kinds of relationships with other constructs. If your hypotheses are supported by the data, then you can have greater faith in both your hypotheses and the validity of your questionnaires. If your questionnaire data do not support your hypotheses, then either your hypotheses are incorrect or your questionnaires are not valid.

Validity checks may be performed in the following circumstances:

- The two constructs may be correlated with one another. For example, you might compare or correlate test scores on your sexuality knowledge test with scores on a general intelligence test. Because they are measuring different content areas, they should not compare or correlate perfectly. However, because they both measure aspects of knowledge, they theoretically should be somewhat related.

- Different groups may be expected to perform differently on your questionnaire. For example, you might hypothesize that freshmen would have less knowledge about sexual activity and be less sexually active than seniors. If your questionnaire measuring knowledge and activity supports this hypothesis, then you have some evidence for the validity of your questionnaire. On the other hand, if your questionnaire did not produce the expected results, then either your hypothesis is incorrect, or your questionnaire is invalid.

- Your experimental group should perform differently on the pretests and the posttests. If your questionnaire finds these results, then you have some evidence for the validity of the questionnaire. If you find no change, then either the program did not perform as you hypothesized, or your questionnaires were invalid.

In order for construct validity to be compelling, your hypotheses should be supported by established theory or data. That is, if others do not find your hypotheses compelling, they will not find this evidence for validity compelling. Moreover, you should not generate one hypothesized finding, learn that it was not supported by the data, create a new hypothesis consistent with the data, and then claim this as evidence for construct validity. The established hypothesis must come first and your test second.

Validity and Criterion-Referenced Measures

If you measure criterion-related validity, predictive validity, or construct validity, your assessment may be based upon correlation coefficients or other statistics that need considerable variation. However, if you are using criterion referenced questionnaires and have little variation in your data, then your correlations may show that the questionnaires are not valid, when in fact they are. The reverse is not true -- if your correlations indicate that questionnaires are valid, then they are. This is the same problem that also occurs when measuring reliability.

Conclusion

Generally, you will find assessing reliability far easier than assessing most kinds of validity. If possible, you should try to obtain evidence for validity. You may, however, choose to do as many others have done and carefully follow the major guidelines for creating reliable and valid questionnaires and then measure their reliability but not their validity. The numerous procedures described in this chapter, especially those on pretesting, will definitely help produce reliable and valid questionnaires.

CHAPTER 6

MEASURING PARTICIPANTS´ ASSESSMENTS OF THE PROGRAM

This chapter discusses issues in designing questionnaires that ask participants to assess both the characteristics of the course and its effects upon themselves. Such questionnaires might be given at any time during or following the program and are particularly useful for formative evaluations. Later chapters discuss issues in designing questionnaires to measure the actual effects of the program -- questionnaires that would be administered in pretests and posttests. This chapter assumes that you have read Chapter 5 on the fundamentals of questionnaire design.

If properly constructed and completed, program assessments can be more sensitive than change data from pretests and posttests and can yield types of data that cannot be obtained from pretests and posttests. If people reflect thoughtfully about a program and its impact, they can consider feelings and reactions to the program that pretests and posttests cannot possibly measure, and they may recognize subtle changes in themselves that pretests and posttests are insufficiently sensitive to capture.

On the other hand, program assessments by participants are also notoriously unreliable and invalid. Participants who like the staff of the program typically rate all the characteristics of the program very positively. Participants who like a course or program invariably indicate that the program had a greater and more positive impact upon them than the program actually had -- especially when the program evaluation is undertaken during the last part of a program when participants may be particularly enthusiastic about the program.

Using Participants´ Assessments

Considering the advantages and disadvantages of program assessments, you can profitably use them on several occasions:

- You want to know the participants´ views on the different parts of the program.
- You want to ask questions (about the staff, particular topics, atmosphere, etc.) that other kinds of questionnaires do not measure.
- You want suggestions for changes and improvements.
- You want to make immediate improvements and cannot wait to compare pretests and posttests.
- You were unable or failed to obtain pretest data.
- You were unable or failed to obtain posttest data.
- You have insufficient resources to do more than a formative assessment of the program.
- You want to use more than one method of evaluating the program. For example, you want to compare self-assessment data with change data from pretests and posttests.

Writing Questions

Nearly all the steps for writing questions that are discussed in Chapter 5 also apply to these questionnaires -- you should specify the important features and outcomes to be measured, construct the questionnaire, pretest it with other experts and participants, and if possible assess its reliability and validity.

Your questions about the features of programs should include questions on the following:

- the program structure (the convenience of the time, duration, place, etc.)
- the staff and their skills
- the topics covered
- the class atmosphere (comfortable, boring, personal, etc.)
- the interaction between the staff and participants.

Your questions about the outcomes of programs should include measures of all outcomes specified as important, if these can be measured by this type of questionnaire. In addition:

- Include open-ended questions about the weaknesses of the program and suggestions for improvement.

- Avoid asking participants questions they cannot answer correctly. Participants may be able to assess how the program has already affected them; some respondents may be able to estimate how the program will affect them in the short term future. However, few people can accurately assess how a program will affect them in later years.

- Allow for negative change as well as positive change. Too commonly evaluators bias their results by phrasing questions so that there can only be improvement. Consider the following example:

  Biased: As a result of this course, how much clearer are your values?

  ___ not at all clearer
  ___ slightly clearer
  ___ somewhat clearer
  ___ much clearer

  Unbiased: As a result of this course, are your values less clear or more clear?

  ___ much less clear
  ___ less clear
  ___ about the same
  ___ more clear
  ___ much more clear

- Use some open-ended questions. For example, you might ask the respondent to describe any other effects of the program or to make other remarks. Often such open-ended questions provide a wealth of insight and good ideas.

Choosing Response Categories

A variety of different kinds of response categories is available to evaluate the features and outcomes of the program. Table 6-1 contains examples of categories that can be used with many different questions. These examples illustrate important characteristics of response categories:

- Use response categories that fit the question.
- Use between four and seven categories. If you use fewer than four, you may lose detail. If you use more than seven, people may have difficulty answering the questions and the additional precision gained may not be real. If the respondents are very young (e.g., 5 to 12 years old), they may have difficulty with more than 3 categories.
- Use categories that allow for both positive and negative change (questions 6, 7, and 8).
- Make the differences between adjacent categories approximately equal.
- If possible, use the same response categories for several questions. The reader will have an easier time.

Table 6-1

Examples of Different Response Categories

To Measure Features of Programs

1. Did the teacher encourage students to ask whatever questions they had?	___ strongly agree ___ agree ___ neutral ___ disagree ___ strongly disagree
2. Did the teacher show respect toward the students?	___ none at all ___ a small amount ___ a medium amount ___ a large amount ___ a great deal
3. Did the students become uncomfortable when sensitive questions or topics were discussed?	___ almost never ___ sometimes ___ about half the time ___ usually ___ almost always
4. Was the teacher enthusiastic about teaching this course?	___ not at all ___ slightly ___ somewhat ___ very

To Measure Outcomes of Programs

5. What is your opinion of the overall program?	___ very poor ___ poor ___ average ___ good ___ excellent
6. Did the course make you less likely or more likely to think seriously before having sex in the future?	___ much less likely ___ less likely ___ no change ___ more likely ___ much more likely
7. Did the course make your social life better or worse?	___ much worse ___ worse ___ no change ___ better ___ much better
8. Because this course do you now respect yourself less or more?	___ a lot less ___ a little less ___ no change ___ a little more ___ a lot more

CHAPTER 7

DESIGNING KNOWLEDGE TESTS

Many people, especially teachers, are accustomed to writing knowledge tests to grade students in the classroom. However, knowledge tests that are used to assess the knowledge of individual students relative to one another may not be the best knowledge tests for assessing the change in knowledge of a group of students over time. Moreover, many knowledge tests are quickly prepared, contain ambiguous questions, and lack a reasonable reliability and validity.

To develop a reliable, valid, and sensitive knowledge test, you should complete the basic steps described in Chapter 5:

- Determine the important areas to be measured.
- Construct the test.
- Pretest it.
- Assess its reliability and validity.

This chapter discusses special issues to consider when applying those basic steps to knowledge tests and assumes that you have carefully read Chapter 5.

Using Existing Knowledge Tests

As discussed in Chapter 5, existing tests can provide a useful pool of questions, many of which may be valid if they came from carefully constructed and tested questionnaires. However, there are only a few tests that have been carefully developed. (References to them are at the end of the chapter, and a knowledge test that proved useful in the evaluation of several different programs is at the end of this volume.) If you do use existing tests, be sure that they focus upon those specific facts and knowledge components that you should be measuring.

Selecting Formats

The many different formats for knowledge questions that are widely used are not equally good. The section below briefly discusses them in order of preference. Of course, not all test designers would rank formats the same way.

Some topics lend themselves to formats that are inappropriate for other topics. Similarly, students vary as to which formats they handle best. Thus, you may want to use a mixture of formats. For example, to test knowledge of body parts, you may want to combine multiple-choice questions with diagrams requiring labeling.

Multiple Choice

Some multiple choice questions include a stem (the first part of an incomplete statement) followed by several possible answers (that complete the statement). Other multiple choice questions include a direct question with several possible answers.

For most purposes, well constructed multiple-choice questions are the best format. They have many advantages:

- Their multiple possible answers reduces the impact of guessing.
- If created properly, they have one and only one correct answer (which may be "None of the Above" or "All of the above."
- Most students are familiar with the format.
- Multiple-choice questions are relatively easy to answer.
- They are easy to score.
- They can cover a wide variety of topics. Contrary to popular opinion, they can also cover major substantive points as well as specific facts.

The main disadvantage with multiple choice questions is that their incorrect answers may inadvertently teach students incorrect information. This problem can be minimized by reviewing the test after its final administration and by emphasizing that multiple choice questions contain answers that are plausible but incorrect. In general, multiple choice questions have few of the disadvantages that characterize other formats.

Alternative Response

An alternative response question is a multiple-choice question with only two possible answers. For students with very limited reading and cognitive skills, the alternative response format may be more valid, because it requires less reading and is less complex.

A disadvantage with this format is that simply by guessing, students will answer correctly half of the questions. This is especially a problem if you wish to rank individual students. It is much less of a problem if you are comparing a large number of students over time, because you will then be comparing mean scores; if the group is sufficiently large, the number of students who guess many questions correctly will be balanced by those who guess many questions incorrectly. At any rate, more questions must be asked to obtain meaningful data.

True/False

True/false questions are best for measuring knowledge of topics that are unequivocally right or wrong. Often they can test knowledge about specific facts with few words and thus require minimal reading time. Moreover, if properly constructed, they are not complex and do not confuse readers.

However, few topics in sexuality are based upon important facts which are clearly right or wrong. This produces several problems. First, creating a statement that is clearly right or wrong is very difficult. You may believe that a statement is clearly correct when in fact it is not. For example, the statement "All people can catch an STD" may be intended as a true statement that counters the common myth that only certain types of people can catch an STD. But the statement is not true for those people who will never have any sexual activity, who live in remote areas of the world that are devoid of STD, or who are involved in mutually monogamous relationships.

Second, people may answer some true/false questions correctly because they are not informed about possible exceptions or factors which make the issue more complex. Some may miss the question because they are informed about the exceptions and greater complexity. Further, in an effort to make true/false questions clearly right or wrong, test designers may add modifiers or phrase a statement in such a manner that someone ignorant about the true answer could guess the correct answer from the language of the question.

Finally, true/false questions have the same problem as alternative choice questions; namely, students will guess half of them correctly. This problem can be reduced by adding a third option, "Don't Know," and/or by telling students that they will be penalized for guessing. The number of incorrect answers could be subtracted from the number of correct answers, while unanswered questions or questions answered "Don't Know" are not counted.

Matching

This format can be an efficient method of asking several questions with similar kinds of answers. Names of body parts, for example, could be matched with their functions. However, for several reasons matching is not recommended as the primary format in a knowledge test. First, it limits the kinds of questions that can be asked. For example, all the answers in the right hand column must have a similar format (one word answers, parts of a diagram). Second, the contents of the right hand column must also be homogeneous. For example, if the left hand column needs a date and there is only one date in the right hand column, then the two can be matched without knowledge of the correct answer. To prevent this, numerous dates must be included and this limits the kinds of questions that can be asked.

Fill-in-the-Blank

This format has two flaws. First, determining whether or not some answers are correct can require experts, which is costly. Second, some answers may be technically correct, but may not be the desired answer. For example, the question "Columbus discovered America in ____?" can be answered with "1492," "a boat," "the Santa Maria," or "a state of desperation." Sometimes the most informed students are more likely to provide nonstandard but correct answers. Thus, the person who scores such tests must be knowledgeable about the material and must be prepared to give credit to nonstandard answers.

Label the Figure

Asking students to label the parts of a figure or diagram may be the only reasonably direct format for assessing certain types of information. Moreover, this format is commonly used by teachers in earlier grades and thus most young people are familiar with it.

Figures have two disadvantages; they can test only one kind of information, and they must be scored by people familiar with the material.

Essay

Although essay questions may be useful in normal classroom use, scoring essay questions from many students on pretests and posttests is too imprecise, too difficult, and too costly. Thus, this format is not recommended for evaluation of knowledge changes.

Selecting the Number of Questions in Each Content Area

After determining both the important content areas to measure in your knowledge test and the question format(s), you need to determine the number of questions to ask in each content area. As a basic rule of thumb, you should ask between three and five multiple choice questions for each content area. However, other rules also apply. If you have defined your content areas very broadly, then you may wish to ask a few more questions in each area. If you have defined them narrowly, you should ask about three questions. If you have defined many content areas, you should reduce the number of questions in each area to avoid making the test too long. If you are using question formats that are easier to answer than multiple choice and in which guessing has a greater impact, then you should ask more questions. For example, true/false questions are easy to answer and guessing plays a more major role; you should include more of them. Kirkpatrick (1981) suggests that one multiple choice question is about equal to three true/false questions. However, you probably do not need to ask as many as 15 true/false questions in any single content area.

Writing Questions

Writing good questions is an art, and consequently increased practice improves that skill. However, there are numerous guidelines that can help people avoid common errors.

Questions in Any Format

- Ask questions about the most important facts, not about trivia.
- Make sure one and only one answer is correct.
- Use a vocabulary that all the students can easily understand (unless, of course, you are testing vocabulary).
- Use simple sentence structure.
- Avoid trick or misleading questions.
- Be sure that questions are independent, that answers to one question do not depend upon correctly answering another question.
- Avoid double negatives.

Multiple Choice Questions

- Include as much of the necessary information in the stem as possible.
- Make the possible answers as short as possible.
- Include three to five possible answers.
- Make all of the possible answers plausible.
- Avoid over-using "All of the above" and "None of the above" because they tend to confuse students.
- Avoid having "All of the above" and "None of the above" be the correct answer more than half the time they are included as possible answers.
- Avoid overly complicated possible answers like "a and c above."

Alternative Response and True/False Questions

- Be sure one answer is clearly better than another.
- Avoid words which tend to make statements false ("all," "always," "none," "never") and words which tend to make statements true ("sometimes," "under some conditions," and "may").
- Limit statements to a single idea.
- Avoid negatives.

Matching Questions

- Keep the content homogeneous.
- Keep the number of items small.
- Have more answer choices than statements unless an answer can be used twice.
- Arrange the answers in a logical manner, if possible.

Reviewing the Sequence of Correct Answers

It is easy to inadvertently have several consecutive questions or a disproportionate number of questions with the same answer. To correct this, list all the answers and then reorder questions, reorder answers within multiple choice questions, and/or modify the questions.

Conducting an Item Analysis

An item analysis can greatly improve the test's quality and can be an important step in its construction. To conduct an item analysis you must administer the test to a group of people. Ideally, the group will include 30 or more people and will resemble the participants in the course both before and after the participants complete the course. Then the analysis has two major steps.

Step 1: Analyze the Distribution of Answers to Each Question

Observe both the level of difficulty of each question and the number of times that each response to a question is chosen as correct. If a question is so easy that most (about 90%) of the people answer it correctly, that question will be useless in measuring change between the pretest and the posttest. Easy questions should be removed or made more difficult, unless you are using criterion-referenced tests.

Conversely, a question that is so difficult that very few people answer it correctly may not help distinguish among students nor measure increases in knowledge over time. If the percentage of people correctly answering a question is close to chance, then that question is too difficult. For example, approximately 50% of students will correctly answer a true/false question simply by guessing; if only 60% of the students can answer it correctly, it is too difficult. Similarly, approximately 25% of students will correctly answer a multiple choice question with four possible answers; if only 35% of the students actually answer it correctly, then it is too difficult. If you are not using a criterion referenced test and if a more reasonable percentage of students will not answer it correctly after the course, then you should remove or modify it.

Although questions that are either too easy or too difficult should sometimes be removed or modified, the entire knowledge test should contain questions with a rather wide range of difficulty. Lord (1952) provided a widely used chart for the ideal level of difficulty of different kinds of questions:

Type of Question	Average Difficulty (Percent Correct)
5-choice multiple choice	70
4-choice multiple choice	74
3-choice multiple choice	77
True/false or alternative response	85
Matching or completion	50

Of course, this chart is only a guide and not a strict criterion that should be rigorously followed. Although all questions should not be far too difficult nor far too easy, questionnaires can deviate significantly from this guide and still be valid.

As indicated above, you should also observe the number of times that each incorrect answer is selected by the students. If students never or rarely choose an incorrect answer in a multiple choice question, then that answer should be replaced by another incorrect but more believable answer. If students too frequently select an incorrect answer, the answer may be unfairly misleading and you should examine it.

These guidelines for finding and removing or modifying questions that are too easy or too difficult are important if you wish either to rank individuals or to measure change over time. However, if you are using criterion referenced methods and if your purpose is to ascertain the percentage of students that meet specified criteria both before and after a program, then it is important NOT to remove or modify items simply because items are too easy or difficult. If you did so, you could destroy the criteria that you or experts carefully constructed and reach incorrect conclusions about the knowledge of the students. For example, if you removed all knowledge questions that everyone answered correctly, then on later administrations of the questionnaire you would incorrectly believe that students were less knowledgeable about that topic than they actually are. Of course, the converse is also true -- if you remove all questions that everyone misses, then later you would incorrectly conclude that the students had performed better than they actually had.

Step 2: Analyze the Reliability of Each Question

If the knowledge test is reliable, students who score higher on the test are more informed about the content areas than students who score lower. Therefore, high scoring students should be more likely to answer any question correctly than should low scoring students. If low scoring students tend to answer a question correctly and high scoring students tend to miss it, then the question may be poorly worded and unreliable or invalid, and you should consider improving it.

There are two ways to assess whether high scoring students are more likely than low scoring students to answer a question correctly. The first method is simpler if you do not have a computer or do not understand correlation. However, it is adequate but not quite as elegant as the second method.

1. Separate the 25% of the questionnaires with the highest total scores and the 25% of the questionnaires with lowest total scores from the remainder of the questionnaires.
2. For each of these two groups of questionnaires, calculate the number of correct responses to each question.
3. For each question, calculate the mean of the high scorers and the mean of the low scorers and compare them.

If more high scorers answered a question correctly than low scorers, then the question is probably reliable. If more low scorers answered a question correctly, then the question may be unreliable and you should examine it and possibly modify or remove it.

The second method is easier and more valid if you have a computer and understand correlation. Correlate the correctness of each question with the overall test score. That is, correlate whether or not question 1 is correct with the total test score; then correlate whether or not question 2 is correct with the overall test score; etc. If the correlation between a specific question and the total test score is high, then students who answered that question correctly also scored well on the entire test, and the question is reliable. If the correlation is low, then the students who answered that question correctly were low scorers on the other questions, and the question is probably unreliable.

References

Existing Knowledge Tests on Sexuality

Allgeir, A.R. Sex Knowledge Survey. Midwestern Psychological Association Meetings, 1978.

Alter, J., & Wilson, P. Teaching Parents to Be the Primary Sexuality Educators of Their Children: Guide to Designing and Implementing Multisession Courses, Atlanta: Centers for Disease Control, 1982.

Clark & Hicks. Sex Information Questionnaire. Atlanta Adolescent Pregnancy Project, 1969.

Kirby, D., & Alter, J. Knowledge Test, An Analysis of U.S. Sex Education Programs and Evaluation Methods, Atlanta: Centers for Disease Control, 1979.

Lief, H.I., & Reed, D.M. Sex Knowledge and Attitude Test. Philadelphia: University of Pennsylvania School of Medicine, 1972.

Petersen, J.C., Ryerson, W., Morris, L.A., & Senderowitz, J. The Sex Attitude and Knowledge Survey, Arizona, Behavior Associates, 1978.

Designing Knowledge Tests

Kirkpatrick, J.S. The Magic of Structure. New York: Planned Parenthood Federation of America, 1981.

Lord, F.M. The relationship of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 1/: 181-194.

Morris, L., & Fitz-Gibbon, C. How to Measure Achievement. Beverly Hills, Calif.: Sage Publications, 1978.

Thorndike, R., & Hagen, E. Measurement and Evaluation in Psychology and Education. Fourth Edition. New York: John Wiley & Sons, 1977.

CHAPTER 8

DESIGNING QUESTIONNAIRES TO MEASURE ATTITUDES, VALUES, AND FEELINGS

Attitudes about sexuality are important for at least two reasons. Some, such as self esteem (positive attitude toward oneself), we value intrinsically. In addition, a person's attitudes may significantly affect how he or she behaves. For example, a person's attitude toward contraception may affect his or her current or subsequent use of contraception.

Many activities in sexuality education classes are directed toward attitudes. Some activities may attempt to promote basic values such as the dignity of all human life or the immorality of using physical force in sexual relations. Others attempt to help students clarify their own and their families' values about such issues as premarital sex and contraception.

If your program has goals involving attitudes, values, or feelings, you should not rely on simply measuring the program's impact upon knowledge or behavior, because changes in knowledge may not lead to expected changes in attitude or behavior. Likewise, changes in behavior may not lead to desired changes in attitude or feelings. Thus, you should measure your program's impact on any attitudes, values, and feelings that your goals address.

You should, however, be cautious. There is a large body of literature that demonstrates that attitudes are poorly related to behavior and that many other factors play a far more important role. Very general attitudes are especially poorly related to behavior; attitudes specific to particular behaviors are more highly related to those behaviors.

To develop reliable, valid, and appropriate scales and questionnaires for measuring attitudes, values, and feelings, you should complete the basic steps described in Chapter 5:

- Determine the important attitudes, values, and feelings to be measured.
- Construct the scales and questionnaire.
- Pretest the questionnaire.
- Assess the reliability and validity of each scale.

This chapter discusses special issues to consider when measuring attitudes and assumes that you have carefully read Chapter 5.

Selecting Important Attitudes and Values to Measure

Commonly when deciding which outcomes to measure, you select those which are both important and have a reasonable chance of occurring. However, when measuring attitudes, you should consider both the values of the community and the needs of the students. Just as some communities are very much opposed to teaching specific values, so others are very much opposed to measuring personal values. Members of the community may feel that measuring certain values may suggest to the students that alternative values are acceptable.

You should also consider the privacy of the students. Before trying to measure sensitive and personal attitudes or values, you should be certain that questionnaires will remain anonymous or confidential.

On the other hand, if people opposing your program have claimed that it destroys cherished values, then you may want to measure these attitudes or values in order to determine whether your program has affected them. In such cases, you would be measuring important values that you hope are not affected by the program, not just values that you hope are affected.

Using Scales Constructed by Others

Psychologists have constructed innumerable scales to measure attitudes and other psychological traits of individuals. Some of these scales, especially those that are more psychological (e.g., self esteem scales) may be useful to you as they exist, or they may evoke creative ideas of your own. However, you should be careful using them.

- Relevancy: Some scales may have titles that appear relevant, but if you examine the actual items or questions, you will find that they do not measure the concepts that you wish to measure.
- Reliability and validity: Look for scales that have reliability and validity coefficients in the .80s or higher.
- Appropriateness: Many scales are designed for adults and are too difficult or inappropriate for young people. Their reliability and validity may be high for a different population, but low for yours. Sometimes you can modify the scales slightly (e.g., lower the vocabulary level) and thereby make them appropriate.

At the end of this chapter are references to potentially useful collections of existing scales.

Selecting the Best Scales

All scales summarize an attitude with a single number or score. Because attitudes frequently range over a broad continuum, a scale must also have a range of possible numbers; otherwise the scale would not be sufficiently precise.

You can obtain these scores by asking a single question or a series of questions. However, there are several reasons to use more than one item and to combine the scores. First, each individual item is likely to measure partly the desired concept and partly other undesired factors. For example, some people tend to agree with items regardless of their content; thus, the first item in Table 8-1 would measure not only students' self esteem, but also their tendency to agree. Other respondents may partially misread an item or focus unnecessarily upon some part of it and answer it differently than you intended. Providing several items and designing the items properly minimize these undesired factors and increase both the reliability and validity.

Table 8-1

A Likert Scale to Measure Self Esteem

1. Overall, I am satisfied with myself.

___Strongly Disagree (1)
___Disagree (2)
___Neutral (3)
___Agree (4)
___Strongly Agree (5)

2. I feel that I have many good personal qualities.

___Strongly Disagree (1)
___Disagree (2)
___Neutral (3)
___Agree (4)
___Strongly Agree (5)

3. I feel I do not have much to be proud of.

___Strongly Disagree (5)
___Disagree (4)
___Neutral (3)
___Agree (2)
___Strongly Agree (1)

4. I wish I had more respect for myself.

___Strongly Disagree (5)
___Disagree (4)
___Neutral (3)
___Agree (2)
___Strongly Agree (1)

5. At times, I think I´m no good at all.

___Strongly Disagree (5)
___Disagree (4)
___Neutral (3)
___Agree (2)
___Strongly Agree (1)

Second, an attitude or feeling may have a number of different aspects, and you may want to summarize them. For example, attitudes toward premarital sex are usually quite complex; most people are not simply for or against premarital sex. Their attitudes depend on the circumstances: age, sex, maturity, degree of closeness to the partner, and other factors. Asking several questions and summarizing them in a single score increases the likelihood of obtaining a reliable and valid score.

If you are measuring a relatively unidimensional attitude, value, or feeling, then you should use about five items in your scale. If you're measuring a more complex attitude, then you should use more items, but rarely more than 12 in a single scale.

Likert Scales

Likert scales are the most popular type of scale to measure attitudes, values, and feelings. Likert scales are based upon the idea that people's attitudes, values, and feelings typically have both a direction (for or against) and an intensity (neutral to strong). Table 8-1 provides an example of a Likert scale that can be used to measure attitude toward self (self esteem). Many other examples are in the Attitude and Value Inventory in the appendix. They all demonstrate a number of important properties of Likert scales.

Disagree-agree responses. All Likert items include a statement and a set of responses that range from strongly disagree to strongly agree. Thus, they measure both the direction and the intensity of the attitude. Sometimes evaluators label the middle category "Undecided" instead of "Neutral" or omit the middle category altogether to force respondents to side either for or against something. In general, you should not force people to make such a choice unless you have a particular reason for doing so.

Strong items. If you use neutral or bland items, everyone may agree and then you have obtained relatively little information. Thus, strongly worded items will be more informative. On the other hand, if you are using criterion referenced methods, strongly worded items may specify a higher criterion than you believe is necessary.

Positive and negative items. In the table, the first two items are positive items; greater agreement with the item means more self esteem. The last three items are negative items; greater agreement with them means less self esteem. When respondents answer numerous consecutive positive questions, they tend to give less attention to each item and to give the same answer to all the questions. These sequences of identical answers are called response sets. In contrast, when some items are positive and others are negative, respondents tend to read them more thoughtfully and to give more accurate and valid responses. Moreover, some respondents have a predisposition to agree or disagree with items. Including both positive and negative items will reduce the effects of this predisposition. In sum, to reduce response sets and to improve the reliability and validity, include both positive and negative items.

Scoring Likert scales. At the right of each response in Table 8-1 is a number in parentheses that represents the score for each response. This score gives information about both the direction and intensity of the attitude. Because the direction of items 1 and 2 is the reverse of the direction of items 3 to 5, the numerical scores assigned to the categories are also reversed (see the numbers in parentheses). With scores on items 3 to 5 reversed, 1 indicates a strong and negative self esteem and 5 represents a strong and positive self esteem. With these scores, you can simply add (or find the mean) of each individual's scores. In this example, a total score of 25 (or a mean of 5) would represent the highest possible self esteem; 20 (or a mean of 4) would still represent high self esteem, but with less intensity; 5 (or a mean of 1) would represent the lowest possible self esteem.

Rating Scales

Rating scales are another popular method of measuring attitudes. Their greatest advantage is their flexibility. You can use each question as a separate scale or combine several questions into multi-item scales. Rating scales can measure attitudes or feelings about a wide variety of phenomena such as dating, premarital sexual behavior, birth control, the opposite sex, interaction with parents, family regulations, clarity of values, and clarity of long term goals. Note that you can use them both to measure the respondents' attitudes toward some phenomenon (the first example below) and to measure the respondents' self assessment of their attitude toward some phenomenon (the second example below).

	Never			About Half the Time			Always
How often should people use birth control if they do not wish to have children at that time?	1	2	3	4	5	6	7

	Not at All Strongly						Very Strongly
How strongly do you feel that people should use birth control if they do not wish to have children at that time?	1	2	3	4	5	6	7

Rating scales have several important traits:

About 5 to 7 categories. If you have fewer than five categories, you may unnecessarily lose detail. If you have more than seven categories, the respondents may have difficulty choosing a category. Furthermore, the apparent additional specificity may be misleading.

Labels at the ends of the continuum. Specify the continuum and tie down the end points by giving the end categories labels. If it is easy to do so, you should also give labels to other categories. Be sure that the apparent distances between adjacent categories and labels are equal. Also be sure the end labels are reasonable enough for some people to choose; that is, avoid extremes that no one will choose.

Scoring. For single-item scales, simply use the numerical scores. If you have multi-item scales, score them the same way you score Likert scales, being careful to reverse the scores of any negative items.

Semantic Differential Scales

Another popular kind of attitude scale, the semantic differential, also involves ratings of feelings about a concept (Table 8-2). Semantic differential scales have the following characteristics:

Table 8-2

Semantic Differential Scale Used to Measure Attitude toward Contraception

Directions: Indicate your feeling toward contraception by reading the pair of words on each line and quickly checking the line that best indicates your feeling. If you have no feelings about contraception, check the middle space. If your feeling is more similar to the word on the left, check a line closer to the word on the left. If it is more similar to the word on the right, check a line closer to the word on the right. Answer once and only once for each pair of words.

good ___ ___ ___ ___ ___ ___ ___ bad

wrong ___ ___ ___ ___ ___ ___ ___ right

responsible ___ ___ ___ ___ ___ ___ ___ irresponsible

fair ___ ___ ___ ___ ___ ___ ___ unfair

strong ___ ___ ___ ___ ___ ___ ___ weak

ineffective ___ ___ ___ ___ ___ ___ ___ effective

cold ___ ___ ___ ___ ___ ___ ___ warm

dirty ___ ___ ___ ___ ___ ___ ___ clean

Adjectives and antonyms. Each scale specifies a particular concept and then contains a series of adjectives and their antonyms that might describe the concept. Any set of adjectives and their antonyms can be used, provided, of course, that they have some relevance to the concept in question. The adjectives and their antonyms are separated by either five or seven underlined spaces.

Random order of adjectives. When creating a scale, randomly order the side on which the positive adjective is presented. For the same reasons discussed about Likert scales, avoid having all positive words on one side and all negative words on the other.

Because the semantic differential forces people to choose between single adjectives, it best measures general impressions rather than specific, complex, or detailed attitudes. Specific attitudes are measured better with a Likert scale.

Scoring. Semantic differential scales may be scored similarly to Likert scales. Assign each line a number, with 1 always representing the most negative attitude and 7 (or 5) representing the most positive attitude. Then add the scores for all the lines, or find the mean of all the scores.

Previous research has demonstrated that many adjectives are highly related to three basic and independent dimensions: potency (how powerful or effective), activity (how active), and evaluation (how good or bad). Consequently, the semantic differential is good for measuring broad, general feelings about something, but is not good for measuring specific attitudes or attitudes that do not primarily involve potency, activity, and/or evaluation.

Behavioral Intentions

Behavioral intentions questions are specifically designed to better predict behavior. Evaluators often attempt to learn how people do or would behave in sexuality-related situations by asking about attitudes. For example, consider this question which measures a general attitude but not an intention:

Two people who do not wish to have children and who have sex should definitely use some form of contraception.	___ Strongly agree ___ Agree ___ Neutral ___ Disagree ___ Strongly disagree

Attitudes, however, do not always accurately predict behavior. Other factors such as norms, habits, peer pressures, economic factors, personality factors, and special circumstances also influence behavior. Thus, evaluators ask questions about what respondents believe they would actually do in a given set of circumstances. Research demonstrates that this type of question better predicts behavior than questions asking about a general attitude. Consider the following example:

	Very Unlikely			About 50-50			Very Likely
If you knew that you were going to have sex this week, how likely is it that you would use some form of contraception?	1	2	3	4	5	6	7

This question is clearly more behaviorally oriented than the earlier one and would probably better predict behavior. Thus, it will probably better evaluate the impact of a program upon behavior than would attitude questions. Remember, however, that even questions about behavioral intentions cannot always accurately predict behavior, especially in the area of sexual activity among adolescents. For example, many adolescents who believe that they will act responsibly when sexually involved do not do so when they actually become involved.

A disadvantage of these questions is that combining several different questions into a single score is sometimes difficult, although you can describe different situations and find the average of these scores.

Other Kinds of Scales

Psychologists have developed several other kinds of scales that have been commonly used: Thurstone scales, Guttman scales, and various sociometric scales. This handbook does not explain them because they are substantially more difficult to develop than the scales described above and thus are not recommended. However, you can read about them in the reference books listed at the end of the chapter.

Constructing and Pretesting the Scales

Reliable and valid scales are surprisingly difficult to construct. A question or item that is very clear to you may have very different meanings for the respondents. Thus, it is especially important to follow the steps described in Chapter 5 for constructing and pretesting questionnaires.

In addition, if you are creating multi-item scales and have a computer on which you can easily obtain correlation coefficients between different items, follow the procedures below. If you cannot easily obtain correlation coefficients, you should seriously consider using existing scales that have been properly validated.

All the items in a scale should measure the same trait and consequently, should be highly correlated with one another.

- Step 1: Create about twice as many items as you need for each scale.
- Step 2: Administer the questionnaire to at least 50 and preferably 100 people.
- Step 3: Calculate the correlations between each pair of items in each scale. For example, if you wrote 20 items for a scale, find the correlations between each of the 20 items and the other 19 items.
- Step 4: Examine the correlations for each scale and throw out those items which are poorly correlated with the other items. Keep only those items that are highly intercorrelated. Be sure that you keep the correct number of items needed for your scale. For example, if you want a scale with 8 items, keep the 8 items that have the highest intercorrelations.
- If you do not have enough items that are highly intercorrelated, then improve the items, add to them, and repeat this entire process.
- Step 5: Review your final selection to make sure it includes a balance of positive and negative items.

This examination of the correlation coefficients can be improved and simplified by employing factor analysis, but you should attempt this only if you already have a basic understanding of factor analysis. Steps 1 and 2 would be the same.

- Step 3: For each scale, one at a time, run a factor analysis on the items. Either limit the number of possible factors to one or allow more factors, but prevent rotation. The items with the highest factor loadings on the first (or only) factor should be the items that best measure the desired attitude.

- Step 4: Throw out those items with the lowest ratings. If several items have similar ratings, you can throw out the poorest items and repeat the factor analysis. If several items still have similar ratings you can employ other criteria for keeping items (e.g., apparent face validity).
- Step 5: Review your final selection of items to make sure it has a balance of positive and negative items.

References

Books Containing Attitude Scales

Bonjean, C.M., Hill, R.J., & McLemore, S. Sociological Measurement: An Inventory of Scales and Indices. San Francisco: Chandler, 1967.

Buros, O. Mental Measurement Yearbook, 6th ed. Highland Park, N.J.: Gryphon Press, 1970.

Chun, K., Cobb, S., & French, J.R.P. Measures of Psychological Assessment. Ann Arbor, Mich.: University of Michigan, Institute for Social Research, 1973.

Miller, D. Handbook of Research Design and Social Measurement. New York: McKay, 1964.

Robinson, J.P., & Shaver, R. Measures of Social Psychological Attitudes. Ann Arbor, Mich.: University of Michigan, Survey Research Center, 1969.

Shaw, M.E., & Wright, J.M. Scales for the Measurement of Attitudes. New York: McGraw Hill, 1967.

Books on Attitude Scaling

Babbie, E.R. The Practice of Social Research. Belmont, Calif.: Wadsworth Publishing, 1975.

Henerson, M.E., Morris, L.L., & Fitz-Gibbon, C.T. How to Measure Attitudes. Beverly Hills, Calif.: Sage Publications, 1978.

Kerlinger, F.N. Foundations of Behavioral Research, 2d ed. New York: Holt, Rinehart, and Winston, 1973.

Miller, D.C. Handbook of Research Design and Social Measurement, 3d ed. New York: David McKay Co., Inc., 1977.

Thorndike, R.L., & Hagen, E.P. Measurement and Evaluation in Psychology and Education, 4th ed. New York: John Wiley & Sons, 1977.

CHAPTER 9

DESIGNING QUESTIONNAIRES TO MEASURE BEHAVIOR AND SKILLS

Many educational programs are ultimately concerned with influencing long term behavior. Courses may focus upon knowledge, but the supporters of the course often hope or believe that improved knowledge will subsequently improve decisionmaking and behavior. Many sexuality education programs have as explicit or implicit long term goals:

- the increase in communication with parents
- the delay of sexual activity
- the reduction of unwanted pregnancy
- the reduction of sexually transmitted diseases,
- more generally, the improvement of social and sexual relationships.

If your program either explicitly or implicitly has goals regarding behaviors, then you should measure its impact upon these behaviors.

Sexuality educators are often concerned with several different components of behavior:

- the amount and/or frequency of individuals' behavior
- the skillfulness or effectiveness of their behavior
- the individuals' feelings about their behavior.

For example, consider communication between two people about sexuality. Your program and your evaluation may be concerned with 1) whether any communication took place, and if it did, how frequently and for how long, 2) whether the participants used important communication skills, and if they did, how effective the communication was, and 3) how comfortable the participants were.

Similarly, consider the use of birth control methods. The program and evaluation may be concerned with 1) how frequently birth control methods were used, 2) how effectively or properly each method was used, and 3) how comfortably and with what feelings participants obtained and used a method.

The previous chapter discussed methods of measuring attitudes and feelings about different behaviors. This chapter focuses upon methods of measuring the amount and the effectiveness of the behaviors. This chapter assumes that you have carefully read Chapter 5 on the fundamentals of questionnaire design and that you are following its major steps for constructing questionnaires:

- Determine the important behaviors to be measured.
- Construct the test.
- Pretest the questionnaire.
- Assess its reliability and validity.

Determining the Important Behaviors to Be Measured

Social and Political Costs and Benefits

Despite the obvious need to evaluate the impact of programs upon behavior, all of us doing research in sexuality education need to weigh carefully the costs and benefits of doing such research. We need to compare the potential costs to the student respondents, to the program, and to the community with the potential usefulness of the information we collect.

Costs to the students. When Kinsey conducted research in sexuality, he took great care to insure that the anonymity of the information he collected was not breached. He interviewed people individually and anonymously, coded the answers directly with a code that few people know, and kept the coded data in a safe that only a few could open. However, sexuality education programs have difficulty maintaining such rigid controls. When students complete questionnaires in classrooms, when questionnaires are collected in class and/or sent through the mails, and when keypunchers keypunch the data, some of these safeguards are lost. Even if you take all reasonable precautions, no set of safeguards is foolproof.

If only a few people conduct research, then the chances of confidential information being released remain small. When many researchers ask thousands of students to complete sensitive questions, then the chances substantially increase. Therefore, when making a decision about whether to ask a sensitive question, seriously consider the cost of some student answering that question honestly, some other students seeing the answer, and the reputation of that person being affected.

Costs to the program and the community. Any question about sexuality may offend some people, but questions about individual sexual behavior are more likely to evoke a negative reaction from parents or concerned community groups. If you obtain appropriate approval from school boards or other boards, if you notify parents and obtain their approval, and if you follow the other steps described in Chapter 11 on administering the questionnaires, then the chances of a negative reaction from parents or the community are greatly reduced. Nevertheless, you should still consider the possibility of negative reaction and 1) be able to justify the need for each question and 2) be sure of strong administrative support based on careful review of the instruments.

Quality of the data. The validity of questions on sensitive sexual behaviors may be lower than the validity of other questions. This, of course, will depend greatly upon the age and experience of the students you're evaluating. When you decide which behaviors to measure, you should consider the quality of the data and the actual benefit of that data to you if it is not valid data.

Aspects of Behavior That Can Be Measured

There are at least three different aspects of any behavior:

- The number of times the student engages in the behavior
- The skill with which the student engages in the behavior
- The comfort level the student feels during the behavior.

Often more than one of these may be important and should be measured.

Constructing the Questionnaire

Confidentiality and Validity of Behavior Questions

A major problem with questions about sexual behavior is that students may not be willing to answer them honestly and thus they may be invalid. However, you can reduce this source of error by following the suggestions below in your questionnaire design and by following the guidelines in Chapter 11 when administering the questionnaire.

Print questions on only one side of the page. This reduces the ease with which other people can see previously answered questions. If you print questions on both sides of the page and if the pages are stapled together, then other people can more easily see previously printed pages facing up.

Bury sensitive questions among other questions. If sensitive questions are grouped separately, then others can more easily see them and may be more tempted to look at someone else's questionnaire. In particular, sensitive questions should not be separated at the bottom of the last page.

Prevent completion time from predicting sexual activity. If a questionnaire has many questions that must be completed only by those who are sexually active, then students may believe that students who finish the questionnaire first are not sexually active and that those who finish last are sexually active. You should prevent this by 1) having all students read and answer all questions or 2) including many insensitive questions in the questionnaire.

Use the random response technique. The random response technique is one method researchers have used to get a better estimate of sensitive behaviors while absolutely assuring anonymity. As an example, suppose you have two identical glasses filled with water to the same level. You know that the temperature of one glass of water is 40 degrees and you wish to know the temperature of the other glass. Rather than measure it directly with a thermometer, you could add the contents of one glass to the other and measure the temperature of the combined contents. If that water is 50 degrees, you could conclude that the water in the unmeasured glass had been approximately 60 degrees.

To use the random response technique to evaluate a sexuality education program, you should give all respondents two questions, a sensitive question and an easy question, that have the same possible responses (yes or no, a Likert scale, etc.). Each respondent then uses a random method (perhaps the flip of a coin) to determine which question to answer. Thus, no one except the respondent can know how any particular respondent answered any sensitive question, because no one knows whether that respondent even answered the sensitive question. For example, you could ask the students to answer one of the following questions yes or no, depending on their toss of the coin:

Heads: Question A: Were you born between January and June?
Tails: Question B: Have you ever had sexual intercourse?

No one knows whether the student who answers yes is talking about a birthdate or a sexual experience. However, the alternative questions were designed so that the frequencies of answers to both the sensitive and the alternative questions could be determined. The researcher, knowing the probability of students getting heads or tails, and the probability of students being born during the first 6 months, can statistically determine the frequency of sexual intercourse in the group.

In the example above,

(1) (2) (3)
(% answering "yes") = (% answering question A) (% born between Jan and June) +

(4) (5)
(% answering question B) (% having had sex)

The answer to term 1 comes from the data. If a coin is flipped to determine whether Question A or Question B was answered, terms 2 and 4 are both .5. If half of all people in your class were born in January through June, then term 3 is .5. With simple algebra, you can find the answer to term 5, which is what you really want to know.

This method requires 1) considerable additional work and 2) rather large sample sizes. However, if you have reason to suspect that respondents are answering dishonestly because of fear of exposure, you may wish to use this method.

Using Other Techniques for Enhancing Validity

Make vocabulary clear and unambiguous. In the area of sexuality, many words are poorly defined and should be either clearly defined or avoided.

Poor: When did you become sexually involved?
Better: When did you first have sexual intercourse?

Use appropriate response categories. If the behavior you're measuring is discrete, then ask for a frequency of that behavior during a specified time period:

How many times did you have sexual intercourse during the last month?

If the behavior is not discrete, then provide appropriate response categories:

When you talk about sexuality with your girl/boyfriend, how often do you listen to her/his feelings?

___ almost never
___ sometimes
___ half the time
___ usually
___ almost always

Include "Does Not Apply" as a response category. Many behavior questions may not apply to all respondents. To avoid confusing the respondents, be sure to include "Does Not Apply" as a possible response in all appropriate questions.

Make items unidimensional. Too frequently researchers include questions that ask about two separate phenomena in one question:

Poor: How often do you have discussions about sexuality with your parents and friends?
Better: How often do you have discussions about sexuality with your parents?
How often do you have discussions about sexuality with your friends?

Write questions that students can answer accurately. Research in many fields strongly indicates that people poorly remember their own past behavior. The more frequent the behavior, the poorer our memory of the frequency. For example, we may easily remember the one time we have gone to Europe, but we may not be able to remember how many movies we went to in the previous year. In general, questions should deal with recent information:

Poor: How many times did you have sexual intercourse during the last year?
Better: How many times did you have sexual intercourse during the last month?

Avoid or modify leading questions. When measuring sexual behavior, you may want to ask many questions with socially desirable answers. You should either rewrite the question so that its social desirability is not so obvious or eliminate the question altogether because it would be invalid.

Poor: Do you ever discuss anything about sex with your parents?
Better: Last month did you discuss sexuality with your parents?

Poor: Last month how many times did you have sexual intercourse without using any form of birth control?
Better: Last month how many times did you have sexual intercourse? How many times did you use some form of birth control?

Allow students to skip irrelevant questions. Many sexual behaviors are hierarchial -- a student who has never kissed someone has probably never engaged in petting; an individual who has never engaged in petting has probably never engaged in intercourse. This hierarchial principle can be used to have respondents skip questions which are inappropriate and may embarrass them.

Question 35: Have you ever kissed a girl/boy?

If no, skip to Question 40.

Question 36: Have you ever had sexual intercourse?

. . . .
. . . .
. . . .

Question 40: How often do you go to the movies?

If you use skip patterns, you risk students' concluding that some students are sexually active (because they completed all the questions and took longer) and others are not (because they skipped many questions and finished quickly). Therefore, use this technique sparingly.

Create separate versions of the questionnaire for males and females. Administering only one version of a questionnaire is often easier than administering two versions. Nevertheless, if you want to ask males and females different questions, or if you want to avoid the awkwardness of terms like "boy/girlfriend," then you may want to consider having male and female versions of the questionnaire. Different versions may also increase confidentiality.

Measuring Skills and Effectiveness

It is often very difficult to measure either skills or the effectiveness of using those skills. In particular, it is difficult to measure communication, decisionmaking, or other interpersonal skills with questionnaires. A number of people have tried and have not been fully successful.

One partially successful approach is to ask questions about frequency: how often has the respondent actually used the various important components of communication, decisionmaking, or other interpersonal skills. For example, to tap decisionmaking behavior, you could ask questions about the frequency with which students consider alternatives, obtain additional information, weigh the outcomes, and take responsibility for the outcomes. This was the approach that we used in our evaluation. A scale based upon this approach is included in the appendix.

Another approach is to focus upon the last event of a particular type and ask numercus questions about that event. For example, if you are trying to measure the effectiveness of using a particular birth control method, you could ask several questions about how respondents used that method on the last occasion. If sufficient time has passed, you can even ask whether the woman became pregnant.

If you are giving questionnaires to a small number of students and can carefully score answers to questions, then you might consider writing scenarios, asking the students to describe the factors they would consider in making a decision, then scoring the answers. Such questionnaires must be carefully pretested and the judges who are doing the scoring must do so blindly; that is, they should not know which questionnaires are pretests and which posttests or which belong to the experimental or control groups.

Pretesting the Questionnaire

When you pretest the questionnaire with a small group of students, you should focus your questions on their perceptions of the sensitivity of the questions:

- Were they comfortable answering the questions?
- Do they think other students would answer the questions honestly?
- Could the questions be reworded so that they would be less sensitive, yet still measure the same behaviors?
- Which questions did they feel were especially likely to elicit incorrect, socially desirable answers?
- How could those questions be reworded?

Assessing Reliability and Validity

The reliability and validity of behavior questions may be reduced by several factors, namely, the respondents':

- fear of being exposed
- reluctance to admit even to themselves that they had engaged in some behaviors
- feelings of guilt about some behavior
- reluctance to remember or focus upon past and painful experiences
- desire to boast and enlarge upon their sexual activities.

86

Thus, assessing reliability and validity of behavior questions is particularly important.

Assessing Reliability

The best method of assessing reliability is the test-retest method. You should administer a questionnaire to the students on two different occasions about 2 weeks apart. Some kinds of sexual behavior are sporadic and change daily or weekly. If students are having sexual intercourse, for example, that behavior probably varies considerably from week to week. To measure the test-retest reliability of questions about sexual intercourse, you should probably administer the questions only a couple of days apart.

Other methods of assessing reliability are not likely to be effective, because you cannot ask several slightly different questions about the same behavior and have the respondents answer them independently. For example, if you ask several slightly different questions, all of which seek to measure the amount of sexual activity the previous week, the respondent will probably recognize their similarity and answer them all the same. Thus, the answers will not be independent measures, and you cannot use split-half or multi-item methods of reliability.

However, you may be able to examine the internal consistency of different questions. For example, you might include the following three questions:

> Have you ever had sexual intercourse?
>
> How many times did you have sexual intercourse during the last month?
>
> If you occasionally have sex, how comfortable are you getting some form of birth control? (Include "Does Not Apply" as a response.)

Numerous combinations of possible answers would not be appropriate. For example, if the respondent said that he had never had sex, but had sex four times last month, and that the last question did not apply, then one or more of his answers must be invalid.

Assessing Validity

As indicated in Chapter 5, assessing the validity of behavior will probably be difficult. Occasionally you can obtain evidence for different kinds of validity.

Face validity. If your questions are truly clear and straightforward, and if the respondents are willing to answer them honestly, the questions may have considerable face validity. However, as discussed in Chapter 5, face validity is the weakest kind of validity and is especially weak if the respondents may be reluctant to answer sensitive questions.

Criterion validity. Criterion validity is the best form of validity, but you cannot obtain it for many behavior questions. Occasionally you can obtain criterion-validity on contraceptive use by obtaining independent data from clinics or other contraceptive sources in the area and comparing this data with the questionnaire data. Sometimes you can also compare student data on communication with parents with parent data on communication with their teenagers. However, if the parents′ data does not support the students′ data, you don′t know whether the parents' data or the students' data or both are invalid.

Construct validity. Often you can hypothesize that different groups of people will engage in different behaviors. For example, freshmen should engage in less sexual behavior than seniors. This enables you to use construct validity in many cases. Unfortunately, freshmen and seniors are very different in many ways and thus the mere fact that freshmen have reported less sexual activity than seniors does not provide good evidence for the fact that you are measuring what you want to be measuring.

CHAPTER 10

SELECTING A SAMPLE

In research, a population is defined as the collection of people (or other phenomena) that are of interest. The first step in evaluating sexuality education or any program is to define the population of interest. If you want to generalize to all teenagers, then the population is teenagers. If you are solely concerned with those teenagers in a particular geographical area or in a particular school, those teenagers form the population of interest.

In many cases, the population is too large to study in its entirety and a portion of the population, called a sample, is studied in order to make inferences or generalizations to the total population. Therefore, the second step is to decide whether to evaluate the entire population or to select a sample. If the population is small -- for example, if a school program has been given to only 100 students -- questionnaires can be given to the whole population. If thousands of students have taken a program, querying the whole population may be too costly or time consuming and you can save time and effort by carefully selecting a sample of students to participate in the evaluation.

Two factors determine the overall quality of the sample: its size and its randomness. Both are important. If the sample is perfectly random but very small, you cannot generalize to a larger population. For example, if 1,000 students participated in a sexuality education program and if you interviewed or gave questionnaires to only 10 of the students, you could not meaningfully generalize to all 1,000 students because the 10 students might have special qualities that make them different.

Similarly, if the sample is very large but not random, the results may not represent the total population. For example, researchers in a major study of sexuality once collected more than 100,000 questionnaires, but their sample was not chosen randomly, so that in spite of their enormous sample size, their results cannot be used to make meaningful inferences to any larger population. A random sample of only 500 respondents would have been more useful.

Selecting a Sample Size

Other things being equal, large samples are better than small samples for two reasons: they decrease the amount of error caused by sampling and they increase the power of the test. Both of these are discussed below. However, they also cost more and may be more difficult to obtain. Thus, when selecting a sample size, you need to consider the amount of acceptable sampling error, the desired power, the feasibility, and the economic and social costs of samples of different sizes.

Sampling Error and Power

Whenever you take a sample, you introduce a certain amount of sampling error. For example, if you toss a fair coin 10 times, you will not always get exactly 5 heads and 5 tails; often you will get 6 heads and 4 tails, 4 heads and 6 tails or some other combination. Similarly if you have 1,000 students in a high school and randomly select 50 for a sexuality education class and another 50 for a control group, the two groups will probably not be identical even before the course begins; one of the groups is likely to be slightly brighter, more sexually active, or be different in some other way. The difference between the two groups is caused by sampling error.

If you select 50 students for both the control group and the experimental group, the difference between the two groups will probably be less than if you selected only 2 students for each group. This illustrates the general principle that larger samples tend to have less sampling error.

Reducing sampling error or its possibility will reduce the probability that you will make either of two errors. By increasing your sample size, you become less likely to erroneously conclude that the difference between an experimental and control group (or between the pretests and posttests) is due to the program when in fact it is due to sampling error. Assume, for example, that the mean number of correct answers on a knowledge test administered after a program is 85 for the experimental group and only 80 for the control group. If there are only 2 students in each of the groups, this difference might have been caused entirely by the differences in the students before the program began. Concluding that the program had caused the difference would be wrong. In contrast, if you had randomly assigned 50 students to each group, this difference is less likely to have been caused by sampling alone, and you are less likely to reach an incorrect conclusion. Thus, with a larger sample, you can be more confident that a difference is actually caused by the program and not by sampling error.

Conversely, increasing the sample size decreases the probability that you will incorrectly decide that a difference between the experimental group and the control group is not significant when in fact it is. In the example above with two students in each group, you might have incorrectly decided that the difference in mean test scores was entirely due to sampling error when in fact it was caused by the program. However, if you had had 50 students in each group, you would probably have accurately concluded that the program was effective. Decreasing the probability of this type of error is called increasing the power of the test.

Statistical principles demonstrate that you can be 95 percent certain that sampling error will be less than or equal to: (1.96)(standard deviation)/(square root of the sample size).

$$\text{Error} \leq \frac{1.96 \text{ standard deviation}}{\sqrt{N}}$$

For example, if the standard deviation is 5 and your sample is 100, your sample estimate of the mean will be within (1.96)(5)/10 or .98 of the true population mean.

Standard deviations are measures of the extent to which the scores are spread out. They are more fully discussed in Chapter 14. However, even when you understand standard deviations, you cannot know what the standard deviation of the data will be until you have collected the data, and of course it is too late to determine the sample size at that time. Thus, you have to make an intelligent guess. You can improve your estimate by observing the standard deviations in previously conducted studies or by asking consultants. One helpful hint: if you have a dichotomous variable that is scored 0 and 1, the maximum possible standard deviation is .5.

Normally, you would choose an acceptable error, and then calculate the needed sample size. Using simple algebra, the formula above becomes:

$$\text{sample size} = ((1.96)(\text{standard deviation})/(\text{acceptable error}))^2$$

That is, you should:

1. Multiply 1.96 times the estimated standard deviation.
2. Divide this product by the acceptable error.
3. Square this quotient.

If you have a dichotomous variable and your estimated standard deviation is .5, and if your acceptable error is .1, then your sample size should be:

$$((1.96)(.5)/(.1))^2 = 96$$

As you can see from this formula, larger samples sizes produce less error than smaller sample sizes. However, increasing sample size carries diminishing returns. Once the sample size is very large, further increases have only a small effect upon sampling error, for the amount of error in the sample statistics is inversely proportional to the square root of the sample size. For example, if you increase the sample size by a factor of 4 (e.g., from 100 respondents to 400), you will reduce the error in sample statistics by a factor of only two.

Although the formula above gives you the proper sample size for a specified amount of error, many evaluators simply use general guidelines for sample size. These are presented below.

Sample Size	Comments
25	This size is about the smallest size that warrants doing statistical research. Effects of the program would have to be rather large to obtain statistically significant results.
100	This size is substantially better than 25 and is commonly worth the additional effort to collect and analyze the data.
200	This size is about the largest that warrants additional effort unless the evaluation is a major project being completed with great care.
1000	This size is necessary for national studies of major significance and requires substantial funding.

Feasibility

Sometimes it is not possible to select a sufficiently large sample; the number of people in the program may be small and thereby limit the sample size; you cannot obtain the names or addresses of previous participants; parents, school boards, or other bodies will not provide their consent; participants will not agree to participate; or other very practical matters will limit the sample. These problems are especially likely to arise when you wish to administer questionnaires to a control group that has not participated in your program.

You need to consider all these potentially limiting factors in advance and surmount them as best as you can. Those factors that can not be surmounted should be described in your final report.

Social and Economic Costs

If the number of program participants and other similar factors do not limit the sample size, then you need to consider the social and financial costs of increasing the sample size. Although the gains from increasing the sample size diminish with increasing sample size, many of the costs increase proportionately. Doubling the sample size may well double the cost of copying the questionnaires, and the person hours required to complete, code, and keypunch the questionnaires. It may also double the risk of students seeing the confidential answers of other students. The person and computer time required to analyze the data will probably also increase, though not proportionately.

In sum, your selection of a sample size should reflect all these factors. Although the optimal sample size will vary from one study to another, the guidelines provided above may be helpful.

Improving the Randomness of a Sample

A sample is considered random if some method that is completely unrelated to any characteristic of the population is used to select the sample. For example, if you listed all the names of the students in the population of interest, selected each name one at a time, and flipped a coin to determine whether to include that person in the sample, you would create a random sample. If the sample is large enough, the students in the sample should have characteristics very similar to the entire population. However, if your population of interest is everyone in a school, and you selected a sample by specifying everyone in study hall, the sample may not be representative, because it would exclude all students who do not come to study hall: students, for example, who are on work programs and students who are preparing for college and do not have time to take study hall. Such students might be affected by a sexuality education program differently than others. Thus, the study hall sample could bias any conclusions drawn about the impact of the program upon the entire school.

There are several methods of randomly selecting students. One of the best is to determine the desired sample size, assign everyone in the population a number, and then using a table of random numbers in a statistics book, select students one at a time until you have the desired sample size. Another good way is to determine the desired sample size, divide the population by that number to determine "p", arrange all names in alphabetical order, and then assign every pth person on the list to the sample. For example, if the population has 2,000 people and you want a sample size of 200 people, you would select every 10th person.

Randomly selecting individuals is often not feasible because it requires calling students out of class. Another method is to administer questionnaires to all the students in each of a random sample of classes in the school. The risk here is the same as in the study hall example above; you must be sure the sample of classes is representative of the total population. The sample should include proportionate numbers of students according to intelligence, racial background, grade level, popularity among peers, etc.

Improving the Response Rates

When researchers try to collect information from a specified sample, they normally are unable to collect the desired information from all the members of that sample. For example, if they mail a questionnaire to a sample of people, the addresses of some of the envelopes may be incorrect. Some of the people may fail to complete and return the questionnaires. Others may not treat the questionnaire seriously, and answer questions in a flippant manner so that the questionnaire must be discarded. The percentage of members of the originally specified sample that provides usable information is defined as the response rate.

You should try to obtain as high a response rate as possible, because a high response rate will help you obtain a sample size closer to your desired sample size. Even more important, a high response rate will help maintain the randomness of your sample. If your response rate is unexpectedly low, you may not have enough people in your sample, and even more important, the people who do not respond may differ significantly from those who do respond and thereby bias your analysis. For example, people who fooled around during the sexuality education course may have learned less and may be less likely to return mailed questionnaires. Thus, if your response rate is low, you might incorrectly conclude that your students learned more than they actually learned. Alternatively, students who were initially less knowledgeable about sexuality may have learned the most, and these students might be less likely to return the questionnaire. In this case, if you got a low response rate, you might incorrectly conclude that your students learned less than they actually did. The important point is that your sample should be a random selection of the people in your population, and if many people fail to return questionnaires, you may not know how this will affect your analysis.

Response rates of 80 or 90 percent are considered very good in social science research. When questionnaires are mailed to people, the response rates are more commonly around about 50 or 60 percent. Responses rates lower than that are generally unacceptable.

When you administer questionnaires to a captive group, you should get usable information from most or all of the group. However, if you send questionnaires home with your students or through the mail to parents or other members of the community, there are several ways to increase a response rate that might otherwise be low:

- Telephone the respondents or send them a notice in advance indicating that they will be receiving a questionnaire in the mail and that it is important for them to complete and return it.
- Offer to share with them or publish the results of the survey.
- Make all correspondence and the questionnaire very professional.
- Ask a principal, school board, or some other respectable person or group to endorse the study.

- Send a followup questionnaire or postcard or both to those who do not respond. (If the questionnaire is anonymous, you must provide a separate, return postcard indicating they have completed and mailed the questionnaire.)
- Telephone those who do not respond.

When your response rates are low, you should determine as best as you can how your respondents differed from the nonrespondents. This is difficult because by definition you do not get the questionnaires from the nonrespondents. However, there are still two approaches you can follow. First, compare the characteristics of your completed sample with characteristics of the population. For example, if most of your completed sample is of one race, but most of the population is of a different race, then there is a bias; or if most of the sample consists of seniors, but most of the students in the course are sophmores and juniors, then there is a bias. Second, make a greater effort to obtain information from some of the nonrespondents and then see if they differ from the respondents. For example, carefully obtain information from about 10 to 20 nonresponding students, and see if they were more likely to have dropped out of school, to have become pregnant, or to have done something else which would bias your data. If they simply moved away because their parents changed employment, then this might not be a significant or important bias.

The Sampling of Programs

When people evaluate sexuality education programs, they typically select first a program, then a sample of participants, and then evaluate the impact of the program upon those participants. If the program is successful, the evaluator probably writes and successfully publishes an article. Others then read the article and conclude that sexuality education is successful. If the program is not successful, the evaluator probably does not write and publish an article to inform others about the program's lack of success. Or, if the evaluator does write the article, journals may be reluctant to publish it. Such inconsistency produces a bias in the literature: only successes are published and read. This bias can be reduced by randomly choosing a sample of programs and then publishing the results of those programs regardless of whether they are found to be successful. Of course, this requires the cooperation of a journal to publish negative results or non-findings.

Reference

Kish, L. Survey Sampling. New York: Wiley, 1965.

CHAPTER 11

ADMINISTERING QUESTIONNAIRES

Few methodology texts provide guides to administering questionnaires, but our experience has clearly demonstrated that poor administration can completely invalidate good questionnaires and destroy the evaluation. Thus, administering questionnaires properly is just as important as constructing them well.

Obtaining Approval

Because sexuality is a sensitive and controversial topic, it is often important to obtain approval to administer the questionnaires from parents, school officials, human subject review boards, and/or other appropriate organizations. There are several reasons to obtain this approval. First, parents should have the right to prevent their children from reading and answering sensitive questions about sexuality. Similarly, school officials and teachers should also have the right to prevent children in their school or classrooms from completing questionnaires on sensitive subjects. Several people who have evaluated sexuality education programs in this country have learned this principle the hard way. They failed to obtain approval from parents and school personnel, and when their work was discovered, the evaluations were scuttled. Obtaining approval in advance would probably have prevented such a drastic consequence.

Second, obtaining approval prior to administering questionnaires can help prevent potential abuse of sensitive data. Very rarely, if ever, do researchers abuse the collection of sensitive data, but obtaining approval may add safeguards the researchers overlooked.

Finally, approval should be obtained in some places because it is required by law or governmental regulations. In California, for example, parental consent must be obtained before sensitive questionnaires are given to students. Similarly, federal regulations require that various kinds of approval if federal funds are used in the evaluation.

In our experience administering thousands of questions to students, 99% of parents gave permission, less than 1% denied their permission, and none complained at any later time. To obtain permission from parents, write them a letter including the following:

- The rationale or need for the information and the study.
- An accurate summary of the questionnaires.
- A statement of approval received from school boards or other official groups.
- A summary of special procedures that will be followed to ensure voluntariness and anonymity.
- A statement of appreciation for parental support.

Experience has indicated that it is not necessary to send the entire questionnaire home to parents. Doing so may needlessly raise questions. On the other hand, if any parents wish to see the questionnaire, they should certainly be given that opportunity.

At the same time, it is essential to provide an accurate summary of the questionnaires. If the questionnaires include questions about sexuality or behavior, the letter to the parents should state this. If the letter to the parents does not accurately describe the questionnaires, parents will have the right to complain. If the letter accurately describes the questionnaires, parents will have no reason to complain and will probably not do so. Some evaluators have made the description more concrete by including an example or two. In sum, if the letter is well written and its information is accurate and valid, then most parents will provide their consent.

Selecting a Test Administrator

Methodologists differ over the question of whether to have a teacher or some other person administer the questionnaires. If teachers administer and see the test, they may either consciously or unconsciously "teach to the test." For example, they might cover some of the facts asked on a knowledge test immediately prior to the test and thereby bias the results of the test. In addition, the teacher's very presence may bias the students, particularly when the students are being asked to rate the teacher or the class. If the study is evaluating several different classes with different teachers, having a single skilled test administrator give all the tests will ensure that the administration is the same in all the classes and avoid the risk that one or more teachers will fail to properly follow the directions, assure anonymity, answer questions about the test appropriately, etc.

On the other hand, there are also reasons for having the teacher administer the tests. First, just as students often treat a regular teacher with more respect than a substitute, some students will treat a questionnaire far more seriously and answer the questions more reliably if the teacher, instead of an unknown person, administers the questionnaire. Second, some students will answer honestly some sensitive questions only when the teacher has gained their trust and is administering the questionnaire. Third, teachers who have considerable knowledge about sexuality may be better able to answer questions about the test than a test administrator who is not knowledgeable about the field. Finally, employing a test administrator each time a questionnaire is being administered may simply be too costly.

In sum, the researcher needs to consider the pros and cons and the individual abilities of both teachers and others to administer the questionnaires. In some situations, a compromise may be the optimal solution. Especially when trust is an issue, the teacher can emphasize the need to complete the questionnaires carefully and explain that responses will be anonymous, and an independent skilled test administrator can then actually administer the questionnaires and ensure that the correct procedures are followed.

Selecting Dates

When employing an experimental or quasi-experimental design, the researcher will typically want to administer the questionnaires at the beginning of the program (pretests), at the end of the program, and possibly several months or even years after the program (posttests).

Timing of pretests is important. Especially for questionnaires containing sensitive questions about behavior, the researcher may want to delay administering the pretest until considerable rapport has been established between the students and the teacher. For example, if the program lasts an entire semester, the researcher might administer the questionnaires at the end of the first week. Such delays are feasible when the program lasts several weeks or longer and when relatively few topics are covered during the first few days.

Similarly, posttests may best be given earlier than the close of the program. For example, if the program ends near the end of the semester or the academic year, then the first posttest should be given about a week before the end. During the last days of a semester or school year, students are excited; they are ready to leave school and have vacation; they have many other tests. Consequently, they are less likely to answer questionnaires carefully and validly. Again, this consideration applies only to long programs.

Ensuring Voluntariness While Encouraging Cooperation

Because of the personal and sensitive nature of some of the questions, students' cooperation in completing questionnaires must be entirely voluntary. Assure them, both in the written questionnaire directions and verbally during the instructions, that they are not required to complete any questions that make them uncomfortable. If the questionnaires are administered as part of a class, emphasize that their decision to skip part or all of the questions will not affect their grades.

On the other hand, the results of the study will be more valid if a large percentage of the selected participants do complete the questionnaires. Thus, you should encourage (but not pressure) students to participate and should stress the importance of their participation in the study. You can do this by emphasizing that their answers may affect future programs.

Ensuring Anonymity

The personal and sensitive nature of the questionnaires requires that they be truly anonymous. No one (including you, the administrators) should know who completed specific questionnaires. To this end, describe all of the following steps before you give students the questionnaires. Discussing the steps before distributing the questionnaires will not only help them follow the steps, but will also assure them that anonymity is being treated very seriously.

- Physically separate the students so that no student can see the responses of any other student. Separating students may require rearranging desks or using a larger room. Neither teachers nor test administrators should walk around the room, if doing so enables them to see the answers of the students.

- Stress to the students that no-one -- neither the student nor anyone else -- should place any identifying information on any of the questionnaires.

- Ask students to use normal lead pencils to complete the questionnaires. Red, green, or purple ink can destroy anonymity of a questionnaire. Supply pencils, if necessary.

- Give all students identical envelopes into which they will place their questionnaires before turning them in. Once questionnaires have been turned in, mix up the envelopes so that no one knows whose questionnaire is on the top or bottom. Alternatively, let them put their questionnaires anywhere in the middle of the pile of questionnaires or have them drop their questionnaires into a large ballot box.

Using Identification Numbers

Whenever you need to match pretests with posttests, you must use some method that enables you to pair each individual's pretest and posttest, yet maintains the confidentiality of the questionnaires. The best and most common method is to assign unique identification (ID) numbers to the students and to their respective questionnaires. There are several different ways of doing this; each has its own advantages and disadvantages.

In the first method, the researcher randomly assigns each student an ID, keeps a list of students' names and their respective ID numbers, and then during both the pretest and posttest, gives each student the questionnaire with his or her ID number written on the questionnaire. This method is relatively simple and works well methodologically. However, the researcher could take a given questionnaire, observe the ID number on it, and then use the list of names and ID numbers to discover which student answered those questions -- thus destroying the assurance of anonymity. Technically, this problem can be overcome by making sure that the people who see the completed questionnaires never have access to the list of names and ID numbers. However, some students may not fully trust this process; they may realize that if the researchers wanted to, they could bring together the list and the questionnaires and destroy their anonymity.

In the second method, students select their own numbers and then put those numbers on each questionnaire that they complete. The administrator can instruct students to use the month and day of their own or a parent's birthday (June 7 would be 0607), the last four digits of their phone number, or the last four digits of their social security number. There are at least three problems with using student-selected numbers. First, the number may not be unique (e.g., two or more students may have the same birth date). This problem can be minimized by further dividing students into reasonably small groups (e.g., Mr. Jones' second period health class), and by using other identifying information on the questionnaires (e.g., the person's sex, age, or handwriting) when duplication of a number does occur. The second problem with this method is that someone may recognize the number. For example, a student may recognize the birth date or phone number of a friend. However, this problem is relatively minor because other students or friends should never see the completed questionnaires, and the researcher who does see them will not know the birth dates, phone numbers, etc. of the students. Moreover, no one will know whether an ID number is a birthday or some other number. The third problem is that students may forget what number they selected (e.g., their own birthday or their mothers' birthday). You can overcome this by specifying yourself what number they should select. Eliminating the student selection of the type of number will, however, slightly increase the chances that someone else will see the number and identify the respondent.

Giving Directions and Answering Questions

All directions should be written on the questionnaire. However, since many students fail to read directions, they should always be paraphrased verbally. If the questionnaires are carefully designed and pretested, then all the directions should be clear, and no students should fail to understand the questions. However, invariably some students become confused about some questions. A student's misreading the directions and answering all questions in an incorrect way will decrease the validity of the data. In general, the test administrator should do whatever will maximize the validity of the data. Of course, the administrator should not answer any knowledge test questions. Moreover, whatever policy the administrator establishes for answering appropriate questions should apply to all administrations of the test. Providing help on the pretest, but not on the posttest, may introduce bias in the analysis.

In addition, we have found that when the teacher stresses the importance of the study and of careful answers, the students do treat the questions seriously; when teachers fail to stress care, some students are careless.

Allowing Sufficient Time

When measuring the effects of sexuality education programs, there is rarely a need to administer timed tests. If students are hurried, or even if they believe that they may be hurried, they may spend less time on each question and answer less carefully. All students should have sufficient time to answer each question carefully.

Attention span varies with different groups of students and with different kinds of questions. When students complete tests for grades, they may be able to concentrate an hour or more, but when they complete questionnaires that have little impact upon them, their attention span is shorter. Attention span is probably greatest for questions about their behavior that they find interesting, shorter for questions about their attitudes, and shortest for questions testing knowledge. If students cannot complete the questionnaires in 20 to 30 minutes, administer them in 2 or more days, if possible.

CHAPTER 12

USING UNOBTRUSIVE MEASURES

This handbook has frequently emphasized that an important principle in methodology is that phenomena -- whether outcomes or programs -- should be measured or evaluated in two or more ways that are maximally different.

Poor: Using two different questionnaires to measure the outcome of a program.
Better: Administering questionnaires and conducting in-depth interviews.

If the maximally different methods all provide evidence for the same conclusion, then you can have much greater faith in that conclusion. Rarely will results from different methods be identical. However, if one of the methods provides evidence for one conclusion and the other method provides evidence for a different and conflicting conclusion, then one or both of the methods must be incorrect, and you cannot have much faith in your conclusion.

Two maximally different kinds of methods are obtrusive and unobtrusive methods. Previous chapters have discussed the use of questionnaires to evaluate programs. Questionnaires are obtrusive because they intrude into the lives of the respondents, requiring their knowledge, their consent, and even their full cooperation. In contrast are unobtrusive methods that do not intrude into the lives of the participants and do not even require their knowledge, consent, or cooperation.

Developing unobtrusive methods often requires great creativity. To demonstrate the wide range of creative possibilities, this chapter will (1) briefly describe several examples of unobtrusive measures used in other fields, and (2) discuss possible uses in the analysis of sexuality education programs.

Using Unobtrusive Measures in Other Fields

Problem: When television first became popular in this country, many people wanted to ascertain the impact of television upon reading habits. To have administered a questionnaire to a random sample of Americans would have been costly, and previous studies indicated that such studies were invalid because many people forget what they have read, and others exaggerate how much they have read.

Solution: The researchers sampled several libraries in the country and observed the changes over time in the number of different kinds of books that were checked out. They also observed the changes over time in book sales.

Problem: A museum wanted to measure the popularity of different exhibits but didn't want to administer questionnaires or directly observe the number of people viewing the exhibits.

Solution: The museum installed an inexpensive tile floor that wore out rather quickly. Every 6 months they measured the thickness of the tiles. If the tiles were thinner than average in front of a particular exhibit, they concluded that more people had walked or stood in front of that exhibit. To assess the popularity of different exhibits with different age groups, they also counted the number of finger prints of varying heights on the glass in front of the exhibits.

Problem: A county in Kentucky abolished the sale or importation of alcoholic beverages in the county and then wanted to know the actual impact upon drinking habits of the population. Obviously people would not have answered questionnaires honestly.

Solution: First, county officials simply observed the change over time in citations for drunk driving. Second, both before and after the abolition of alcoholic beverages, they counted the number of empty bottles of different kinds of alcohol in a random sample of trash cans in the county.

Using Unobtrusive Methods to Measure Contraceptive, Pregnancy, and STD Rates

Many programs want to 1) increase the effective use of birth control methods, 2) decrease the amount of unprotected sexual activity (and thereby decrease unwanted pregnancies), and 3) reduce the number of cases of sexually transmitted disease. These goals are clearly not the only goals of sexuality education programs. However, many programs consider them very important and use them to justify their funding. Thus, it is essential to adequately measure the impact of programs upon pregnancy and STD rates.

Whereas previous chapters have discussed methods of measuring these rates with questionnaires, this chapter focuses upon unobtrusive methods of measuring these rates, namely, collecting data from clinics.

Collecting Data from School Clinics

In a few high schools, measuring the impact of a program upon contraceptive use, pregnancies, births, and STD's is relatively easy, because most girls who want some form of contraception, who become pregnant, or who get a sexually transmitted disease go to the high school clinic for initial treatment and/or referral. Thus, you can ask the health clinics in these schools to simply tally the number of observed pregnancies, births, and cases of sexually transmitted diseases each year.

In many schools, a large proportion of girls who become pregnant go to term and are either visibly pregnant while at school and/or obtain a medical excuse from the school clinic to drop out of school before and after delivery. In these schools, the health clinics can tally the number of births.

Because such clinics also have access to the names of the students who attend the sexuality education programs, they would be able to 1) determine the numbers of pregnancies, births, and cases of STD each year before and after the sexuality education program and 2) compare people who did and did not take sexuality education. That is, you can use a quasi-experimental design in such settings to obtain important information unobtrusively.

However, monitoring the use of contraception, the number of pregnancies, and the number of cases of STD is much more difficult than monitoring births. Many girls may use contraception, have early miscarriages or abortions, or have STD's without notifying any staff person in the school. Thus, you will normally have difficulty obtaining valid data from school clinics.

Confidentiality is always a serious problem, because a teenager's use of contraceptives, pregnancy, or case of STD should never be made public. To assure that confidentiality is maintained, only appropriate people should be allowed to view the clinics' records, and all research data with personal identifying information should be kept absolutely confidential.

Collecting Data from Nonschool Health Clinics

In some communities the vast majority of teenagers who obtain a medical form of contraception, who become pregnant, or who get an STD go to a small number of doctors or clinics. You can determine if the students in a particular school visit a limited number of doctors or clinics by administering an anonymous questionnaire to seniors, asking them where they have gone or would go if the need should arise. If most students would visit a limited number of doctors or clinics, and if all of these doctors and clinics are willing to participate, then there are two different ways to collect contraceptive, pregnancy, and STD data.

One method involves creating lists of all female students in your school each year, then looking up each student's name in the files of each clinic or doctor to determine whether that female student obtained contraception, got pregnant, or got an STD that year. Often, when looking up a name, it is convenient to see whether that student attended the clinic or doctor during any previous year. When checking the records, write down the date of the visit, so that 1) additional pregnancies or cases of STD by the same person can be recorded, and 2) redundant visits to two or more doctors or clinics for the same problem will not be counted twice. Count the numbers of people who obtained contraceptives, got pregnant, or had an STD, and add across clinics and doctors. Then compare rates before the implementation of a sexuality education program with rates after its implementation.

If you wish to compare students who take sexuality education with other students who have not taken sexuality education, divide your list of female students accordingly. Then find the rates for each group.

If a doctor or clinic from whom you need data will not allow an outsider to view their records, you may be able to hire a staff person currently working for that doctor or clinic to review the records. This procedure may also help maintain the confidentiality of the data.

A second method of collecting contraceptive, pregnancy, or STD data involves asking the cooperating doctors and clinics to collect a very small amount of additional information during the intake interview from all teenagers getting contraception, having a positive pregnancy test, or having an STD. If the doctors and clinics ask which school the teenager attends, they can provide you with the numbers of teenagers from each school each month or year that obtained contraception, had a positive pregnancy test, or had an STD. In addition, if the doctors and clinics ask whether the teenager completed the sexuality education course in that school, then they can provide data on the numbers of teenagers both taking and not taking sexuality education who are seeking treatment or contraceptives.

The first method discussed above has two major advantages. First, it can be implemented months or even years after people have attended the clinic, provided the records of the clinic patients remain on file. In contrast, the second method does not allow the collection of data for years prior to the data collection, because it depends upon patients answering questions when they come to the clinic. Second, the first method is probably more reliable. If you can gain access to the records, and if you have sufficient time to look up all the names, it should be a relatively straightforward task producing reliable data. However, if you are evaluating several schools, or if you are evaluating schools with very large numbers of students, the first method may be too time consuming and costly.

Collecting Data from District or County Statistics

In a few schools you can use district or county statistics to evaluate your program. If a school district implementing a sexuality education program is congruent with a county or a health district, then you may be able to obtain official estimates of contraceptive use, pregnancies, abortions, births, and cases of STD from the county or health organizations. That is, other people will already have done all the work for you. However, you can only use this data to compare the statistics before the sexuality education program was implemented with the statistics after the program was implemented, and rarely are sexuality education programs implemented quickly in entire health districts.

Limitations on Pregnancy, Birth, and STD Data

Because pregnancy, birth, and STD rates vary substantially in schools from year to year simply because of chance factors, obtaining several years of baseline data and several years of post-program data is critical, particularly when the impact of the sexuality education program is likely to be small. For example, if a course successfully reduced the number of teenage pregnancies among the students in the course by 30%, and if 30% of the student body completed the course, then the pregnancy rate for the entire school would decline by only 9% (30% x 30%) because of the program. The rather small impact of a rather successful program would probably be obscured by the annual changes in pregnancy rates caused by chance or systematic factors.

Any procedure for collecting pregnancy, birth, and STD data will invariably underestimate the actual number of pregnancies, births, and cases of STD because some pregnancies, births, and cases of STD will certainly be missed. However, this is not a problem if 1) the researcher is comparing rates over time, and 2) the percentage of missed pregnancies, births, or cases of STD remains constant over time. For example, if a school actually has 100 pregnancies per year before a program but only identifies 80% or 80 of them, and if the same school actually has 70 pregnancies after a program but only identifies 80% or 56 of them, then the researcher would properly conclude that pregnancies have declined by 30%. In sum, a systematic error in the collection of data will not affect the estimated percentage change.

When comparing the contraceptive, pregnancy, and STD rates of students who have taken sexuality education and students who have not taken sexuality education, you should be certain that the two groups are similar in other respects. For example, it would be completely invalid to compare all students who have taken sexuality education with all who have not taken sexuality education, because those who have taken sexuality education are probably older and more sexually experienced. Thus, for example, you should compare freshmen who have taken sexuality education with freshmen who have not; sophomores who have taken sexuality education with sophomores who have not; etc. Similarly, you should control for other important factors that might exist such as race, sex, and religion.

Using Unobtrusive Methods to Evaluate Other Goals

Comprehensive sexuality education programs frequently have many goals other than increasing use of contraception and reducing unwanted pregnancies and cases of STD, and some of these can be measured by unobtrusive methods. If a goal of your program is to increase serious discussion of sexuality and decrease exploitive thinking, you could count both before and after a program is implemented, the number of sexist or "dirty" comments or jokes in the locker rooms of the school, in the hallways, or on the bathroom walls. If a goal is to improve skills in communication and conflict resolution, you could tally the number of fights on the school grounds before and after program implementation. If a goal is to reduce inappropriate public sexual behavior, you could observe any changes in the amount of necking in the hallways. If a goal is to improve thoughtful communication about sexuality, you could monitor the number of articles in the school newspaper, the number of discussion groups about sexuality, or the treatment of sexuality in other school functions.

None of these unobtrusive methods is optimal or ideal. Each of them has one or more problems that may reduce its validity. However, unobtrusive methods can be very useful as methods that are maximally different from obtrusive methods and thus provide an independent source of evidence for the success of a program. You should use your creativity and design additional unobtrusive methods suitable for the evaluation of your own program.

Reference

Webb, E.J., & Campbell, D.T. Nonreactive Measures in the Social Sciences. 2d ed. Boston: Houghton Mifflin, 1981.

CHAPTER 13

PREPARING DATA FOR ANALYSIS

This chapter discusses procedures for preparing questionnaire data for statistical analysis. The next chapter discusses specific statistics to use when analyzing the data.

Doing the Analysis by Hand Versus Computer

When analyzing quantitative data, first decide whether to analyze the data by hand or to use a computer. Doing it by hand is probably best whenever all of the following conditions exist:

- The data contain a small number of cases (e.g., respondents).
- The data contain a small number of variables per case.
- You desire only simple statistical analyses such as frequencies and mean scores.

Using a computer is usually necessary whenever 1) the data contain a large number of cases or variables, or 2) you wish to use more complex statistical analyses such as tests of significance, correlation coefficients, or reliability coefficients. If you decide the data or data analysis require a computer, but do not know how to use one, you can usually hire a university graduate student or other consultant to conduct the statistical analysis. If you do hire someone, make sure that that person has had considerable experience with the kinds of data and data analysis that you will have.

Coding Questionnaire Data

Coding the questionnaire data is the process of translating answers on the questionnaires into numbers (or occasionally letters) that can be subsequently keypunched. Coding can take a substantial amount of time and may introduce additional errors. Therefore, whenever possible, you should design the questionnaires so that the keypuncher can punch the data directly from the questionnaire without someone coding all the data on separate sheets of paper.

Develop a Codebook

The first step in coding the data is to develop a codebook, the directions that translate a set of questions into a matrix of numbers or letters that are to be entered into the computer. The codebook must accurately specify 1) how each answer on the questionnaire should be translated into a letter or number, and 2) the correct column placement in the data matrix for each question or variable.

- If possible, use only numbers in the data matrix.

- Assign each questionnaire a separate ID number. It is often convenient to use sequential numbers, or if you are matching pretests and posttests, to use as the ID number the birthdate or other number that is your basis for matching.

- Code important information that may not be written on the questionnaire: whether the questionnaire is a pretest or posttest, whether the respondent is part of the experimental or control group, which class and teacher the respondent had, etc.

- Be sure that every answer can be assigned one and only one number.

- If respondents write in answers different from the possible choices that you provided in the questionnaire, and if you wish to code these answers, then 1) give each of them a different code, divide the answers into groups, and give each group a separate code, or 2) combine them all into an "Other" category.

- When you have missing data, use a missing data code equal to some number that can never be a valid answer. Coders often use ´9´ as the missing data code for one column answers and ´99´ as the missing data code for two column answers. Other kinds of missing data such as "Does Not Apply" can be given the same missing data code, or if you might want to analyze them separately, a different number such as ´8´ or ´98´ that cannot be valid.

- If feasible, assign columns in the the data matrix in the same order as the questions in the questionnaire.

- Plan coding according to the capability of your equipment. If you will be keypunching onto IBM cards, you will be limited to 80 columns per line. If you will be keypunching directly onto magnetic tape or into the computer, you should probably not exceed 132 columns per line, because most printers which will print out your data file are limited to 132 columns per line.

Code the Data

Coding the data is a process of preparing data so that someone can easily keypunch it. You should either 1) add identification numbers and/or other numbers to the questionnaires so that the keypuncher can keypunch directly from the questionnaire or 2) copy the identification numbers and all other data from the questionnaires onto a specially prepared sheet of paper from which the keypuncher will keypunch the data. If the questionnaire is rather straightforward, then the first way is certainly easier and probably more reliable because it eliminates the tedious step of copying numbers by hand. On the other hand, if the questionnaire is complex and requires some thought to code the questions, then the second method may be necessary.

To include the answers to open-ended questions in the data, you must code them first before they can be keypunched.

When coding any data, be sure to do it carefully. If your concentration diminishes, take breaks.

106

Check the Reliability of the Coding

Whenever the questionnaires include open-ended questions that require careful consideration in coding, at least two people should code some of the questionnaires and their codes should be compared. If the questionnaires include only closed-ended questions that can be coded with little thought, then only one person needs to code the data, but that coder should still periodically complete spot checks of the coding to assure that he or she is not making any errors.

Keypunching the Data

If you are not an experienced keypuncher and have a substantial amount of data, you should probably use a professional keypunching firm. Such firms commonly keypunch data very rapidly and hence also inexpensively; they make far fewer errors than novices; and they utilize computer programs that further reduce errors. For example, some programs check each data entry to ensure that it is an acceptable entry for that column.

If funds are available, verifying the data is usually a relatively inexpensive option that further reduces error. During the process of verification, one or two keypunchers punch each questionnaire twice, and the computer compares the two copies. If there is any discrepancy, the computer alerts the keypuncher who then ascertains the correct entry.

Experience has clearly demonstrated that nonprofessionals make many errors. Thus, if you are not an experienced keypuncher, but nevertheless must do the keypunching, you should definitely use some reliable method for checking errors, such as checking all your work, or preferably, asking someone else to check it.

If the data is not keypunched directly into the computer where it will be analyzed, then it must be put onto some kind of device so that it can be transferred to the desired computer. If the amount of data is relatively small, IBM cards may be easier. If the amount of data is large, then magnetic tapes should be used. Because the specifications for these tapes differ with each computer installation, you should check with the installation.

Setting up Keypunched Data on the Computer

When you put the keypunched data into the computer, the resulting file will automatically be a rectangular matrix of numbers. Sometimes this matrix of numbers can be statistically analyzed by the computer as it is.

However, if you used an experimental design and have both pretests and posttests, then you must treat the data as data from either independent samples or matched samples. Treating the data as data from matched samples allows you to later conduct certain statistical tests that are more powerful and provide more valid information than others that do not require matched samples.

Independent samples. If you do not have the identification numbers for each respondent and cannot or will not match each pretest with each posttest, then you must treat the data as data from independent samples. In this case the pretests and posttests can be entered in any order, although often it is more convenient to enter all the pretests first and all the posttests second. When specifying the names of the variables in the SPSS file (discussed below), use the same names for both pretest and posttest variables. For example, if you enter the answer to question #1 as Q1 on the pretest, you should also call it Q1 on the posttest. You must also have some variable which indicates whether a given case is a pretest or a posttest.

Matched samples. If you have identification numbers and can match each pretest with each posttest, then you should (but do not have to) treat the data as data from matched samples. To do this, you must include for each respondent first the pretest data and then the posttest. To group the data in this manner, you can 1) sort the data in the column by indicating whether the case is a pretest or posttest, 2) sort the data by identification number, and 3) eliminate any cases for which you do not have both pretest and posttest data. If you have to eliminate many cases, then your remaining sample may be biased and you should either treat the data as data from independent samples or treat the data first as data from independent samples and then as data from matched samples and compare the results of the two analyses.

Specify different names for the pretest and posttest variables. For convenience, begin each variable on the pretest with the letter A (A1, A2, A3, etc.) and each variable on the posttest with the letter B (B1, B2, B3, etc.). Any subsequent posttests may be labeled C, D, etc.

Creating an SPSS Program File

Although many computer software packages are available, most researchers who analyze social science data use the Statistical Package for the Social Sciences (SPSS), and it is highly recommended. SPSS can conduct all the kinds of statistical analyses that the user may desire. It can handle missing data (which you are likely to have), and it has excellent handbooks on its use.

Following are several suggestions for setting up the SPSS file cards:

- Assign each variable a label similar to the questionnaire numbers (e.g., Question #1 would be Q1; Question #2 would be Q2; etc.) If the file contains pretest and posttest data, use A1, A2, etc., and B1, B2, etc., as discussed above.

- Use the COUNT card to score knowledge tests.

- Use the RECODE cards to reverse the order of any variables that should be reversed (e.g., negative statements that are part of indices).

- Use the COMPUTE cards to add together different variables to make indices. When doing this, be sure to divide by the number of variables in each index and be sure to use the ASSIGN MISSING card to handle missing data.

Cleaning the Data

Regardless of the care with which students complete questionnaires, coders code them, and keypunchers keypunch them. All these people are human and invariably make mistakes. Some simple coding and keypunching errors can produce incredible errors in conclusions. For example, consider a question asking for the number of times respondents had sex in the last month. Imagine that "99" is being used as the missing data code for a two column answer, and the keypuncher accidentally punches "98" instead; you might improperly conclude that there had been a great change in sexual behavior. Thus, properly checking and cleaning the data can be extremely

108

important and should be completed before conducting any serious analysis.

Scan the Data Matrix

Print out the data file and examine it in the following ways:

- Check to see that all the cases have the correct number of lines.
- If there are pretests and posttests for each case, be sure each case has the required number of tests.
- Be sure all the lines in the files have the correct number of columns. If all the cases have only line, then the right hand boundary should be straight. If the cases have more than one line, then the right hand boundary should be consistent; it should form either a straight line or a regular pattern.
- Be sure other columns appear to be aligned properly. For example, if one or more columns should be blank, scan down the file and be sure that they are in fact blank. Or, if one or more columns should have only "0"s or "1"s in them, be sure that they do in fact contain only "0"s and "1"s.
- Check each case for large amounts of missing data. If a case is missing only a small amount of data, you should keep the case and use one of several useful options SPSS offers to handle missing data. If the case is missing a large amount of data, you should probably exclude the entire case. The missing data may indicate that the respondent had too little time, was confused, or did not treat the questionnaire seriously; consequently, what data does exist may not be valid.
- Check each case for response sets. If you find the same answer for several consecutive questions on a knowledge test or attitude inventory, you should consider excluding the case. The respondent probably did not read these questions or may not have taken the questionnaire seriously. For example, if a respondent answered eight successive questions on an attitude inventory with a "5," and if logically consistent answers would require some high and some low answers, then that data is probably invalid.

When making decisions about whether to exclude cases because of missing data, response sets, or any other reason, you should always do so blindly. That is, you should never determine whether keeping or excluding a case will support or refute your theoretical conclusions and then keep or exclude the case for that reason. To do so would greatly bias your data and would violate basic principles of scientific research. On the other hand, you should not keep data or cases that are clearly invalid, because invalid data can also bias results. Thus, you should exclude cases when the data is clearly unreliable or invalid, but your decisions should be determined entirely by the validity of the data, not by the effect the data would have on your conclusions.

Examine the Frequency Distributions of Each Variable

The FREQUENCIES program in SPSS can provide you with the frequency distribution for each variable. Examine the frequencies to make sure that all the answers that actually occur in the FREQUENCIES printout are reasonable numbers. If some numbers are implausible, look first for the implausible numbers in the data file.

- If the numbers in the data file are not plausible, then return to the original questionnaires, find the error, and correct the data file.

- If the original numbers reported in the questionnaire are excessively large and clearly invalid, simply declare the offending numbers as missing data. For example, if a student claims to have had sex 100 times in the last week, this number should be declared missing data.

- If the numbers are correct in the data file, then the SPSS file must contain errors, and you should correct these.

- If the SPSS file is correct, and it is impossible to find the original correct numbers, simply declare the offending numbers as missing data.

Reference

Nie, N., Hull, C.H., Jenkins, J., Steinbrenner, K., & Bent, D. Statistical Package for the Social Sciences, 2d ed. New York: McGraw-Hill, 1975.

CHAPTER 14

STATISTICAL ANALYSIS

Some people are frightened by statistics, probably because they do not understand them. However, there is nothing inherently magical, mystical, or difficult about statistics. They are simply procedures to summarize data, make the data more clear and understandable, and help answer specific questions.

This chapter discusses the basic procedures of statistical analysis that you will need to analyze your data. It will assume that the computer will do most of the calculating. If you need to know how to calculate by hand any of these procedures, read one of the statistics books referenced at the end of the chapter.

The decision about what statistical procedures to use depends upon the type of data that you have and upon the hypotheses that you wish to test. There are four basic kinds of data: nominal, ordinal, interval, and ratio.

Kinds of Data

Nominal

Nominal variables are characterized by:

- mutually exclusive categories; that is, no person could fit into two or more categories of the same variable.
- exhaustive categories; every person can fit into one category in each variable.
- categories with assigned numbers that have no real meaning apart from their function as codes; a larger number does not mean more or less of something than a smaller number.

The following variables are examples of nominal variables:

Religion:	Race:	Sex:
1=Catholic	1=White	1=female
2=Protestant	2=Black	2=male
3=Jewish	3=Hispanic	
4=Agnostic & Atheist	4=Oriental	
5=Other	5=American Indian	
	6=Other	

Note that you cannot meaningfully add, subtract, divide, or multiply nominal numbers. The average of "1" and "5" is "3," but the average of "Catholic" and "Other" is not "Jewish."

Ordinal

Ordinal variables are characterized by categories that:

- are both mutually exclusive
- have assigned numbers whose order has real meaning; that is, the numbers reflect the order of some real and underlying dimension.

In the example below of political beliefs, people who have a larger number are more "liberal" or "leftist" than people with a lower number.

Political beliefs:
1=radical right
2=conservative
3=middle of the road
4=liberal
5=radical left

Belief about birth control: Two people should definitely use birth control if they are having sex and do not wish to have children:

1=strongly agree
2=agree
3=neutral
4=disagree
5=strongly disagree

Sexual beliefs:
1=intercourse only acceptable in marriage
2=intercourse only acceptable if engaged
3=intercourse only acceptable if in love
4=intercourse only acceptable if there is caring
5=intercourse acceptable any time

Methodologists differ over whether or not you can add and subtract ordinal variables. Methodological purists claim that you do not know that distances between adjacent categories are equal, and thus you cannot add or subtract. Other methodologists claim that commonly the distances are approximately equal, that small errors make little difference, and that the advantages of being able to add and subtract are great. A reasonable criterion is the following: if you believe that you and other methodologists could reasonably consider the distances between adjacent categories about equal, then treat the variable as interval (see discussion below) and add and subtract them. Otherwise, you should not add or subtract them. The first two examples of ordinal variables above could be treated as interval because the categories are approximately equally far apart. The distances between categories in the third example are less certain, but could probably also be treated as interval.

Clearly you cannot multiply and divide the categories. For example, 4 is twice as much as 2, but it would be meaningless to say that a liberal is twice as much as a conservative.

112

Interval

Interval variables have

- mutually exclusive and exhaustive categories
- categories that have a natural order
- categories with meaningful distances between them.

For example, temperature as measured by Fahrenheit or Centigrade is an interval variable; the difference between 70 and 80 degrees is meaningfully the same as the difference between 90 and 100 degrees.

Interval data do lend themselves to some arithmetic operations. You can add and subtract scores meaningfully and multiply or divide the sum or difference of two or more scores. For example, you can average 70 and 80 degrees by adding the two temperatures together (70 + 80 = 150) and dividing the sum by two (150/2 = 75).

Thus, you can calculate means on interval data. However, you cannot meaningfully multiply or divide the original scores; 40 degrees is not twice as hot as 20 degrees.

Surprisingly, you can assign the values 0 and 1 to a dichotomous variable and treat it as interval.

Examples:

0=participated in the control group
1=participated in the experimental group

0=never had sexual intercourse
1=had sexual intercourse

0=had never gone to a clinic for contraception
1=had gone to a clinic for contraception

0=male
1=female

To be interval, the distances between categories must be equal; dichotomous variables meet this criterion because the variables have only one "distance," and thus the distances cannot be unequal. The real reason for treating dichotomous variables as interval is that you can then perform various important arithmetic operations without violating important statistical assumptions and producing nonsensical outcomes.

Ratio

Ratio variables have

- mutually exclusive and exhaustive categories
- categories that have a meaningful order
- distances between the categories that are meaningful
- a meaningful zero point.

The following examples are ratio variables:

- weight
- height
- number of questions answered correctly on a knowledge test
- number of hours that a sexuality education class lasts
- number of times that respondents talked with their parents
- number of times that respondents had sex in the previous month

Because these variables have a meaningful zero point, the ratio of two numbers is meaningful. For example, 80 pounds is twice as much as 40 pounds, and eight acts of intercourse is four times as much as two acts. Therefore, you can legitimately perform all arithmetic operations on the scores. You can add, subtract, multiply, and divide them. This gives them a great advantage over other kinds of variables.

Descriptive Statistics

There are two basic kinds of statistics, descriptive and inferential. Descriptive statistics simply summarize and describe different properties of the data. To make a statement about the students who participated in a program, you would use only descriptive statistics. In contrast, inferential statistics help you make inferences or generalizations from the sample before you to some larger population. If you want to make a statement not only about the respondents, but also about similar students who might take the course, then you should use inferential statistics.

This section will discuss in order the steps that you should follow to simplify and summarize your data. It will use the same data in each step as an example. Suppose that you administered a 40-item knowledge test to 25 students both before and after your course, and that you obtained the test scores in Figure 14-1.

Figure 14-1

Knowledge Test Scores Presented as Original Raw Data

Pretest Scores

31	19	24	28	14	38	21	26	29	12	26	28	37
20	30	24	25	9	30	20	23	28	17	28	34	

Posttest Scores

22	38	28	34	18	36	33	33	24	40	15	27	30
31	40	33	13	21	35	26	38	39	20	34	35	

Arrays

To create an array, simply put all the numbers in numerical order. A computer can do this for you, or you can create one yourself. The scores in the example are presented in Figure 14-2.

Figure 14-2

Pretest and Posttest Scores Ordered in Arrays

Pretest Scores

9	12	14	17	19	20	20	21	23	24	24	25	26
26	28	28	28	28	29	30	30	31	34	37	38	

Posttest Scores

13	15	18	20	21	22	24	26	27	28	30	31	33
33	33	34	34	35	35	36	38	38	39	40	40	

By simply ordering the scores in this example, you can see that posttest scores are larger than the pretest scores. However, if you have a large number of cases, simply presenting the array of scores requires too much space, and you need to summarize further.

Frequency Distributions

A frequency distribution is a table that specifies the number of times that each number appears in the array. In Figure 14-3 are the frequency distributions for the scores in the example. As you can see, they present the scores in a manner that is more clear, understandable, and concise than the original arrays. Note that no information was lost by presenting the scores in this format.

Percentage and Cumulative Percentage Distributions

Often you will find it helpful to know not only how many individuals obtained the different scores, but also the percentage of all individuals who obtained that score. To create the percentage distribution, divide the frequency of each value by the total number of cases. Sometimes you will want to know the percentage of all scores of a particular size or smaller; to find the cumulative percentage, add the percentage to all the percentages of lower scores. Figure 14-4 illustrates these two distributions, using data from the example.

Cumulative percentage distributions have special importance when using criterion referenced methods. Cumulative percentage distributions can show the percentages of students who reached the criterion levels specified previously by experts. You can use them to compare students' performance on the pretest and posttest.

Figure 14-3

Knowledge Test Scores Presented in Frequency Distributions

Pretest Scores (N=25)		Posttest Scores (N=25)	
Value	Frequency	Value	Frequency
9	1	13	1
12	1	15	1
14	1	18	1
17	1	20	1
19	1	21	1
20	2	22	1
21	1	24	1
23	1	26	1
24	2	27	1
25	1	28	1
26	2	30	1
28	4	31	1
29	1	33	3
30	2	34	2
31	1	35	2
34	1	36	1
37	1	38	2
38	1	39	1
		40	2

Figure 14-4

Knowledge Test Scores Presented as Percentage and Cumulative Percentage Distributions

Pretest Scores
(N=25)

Value	Frequency	Percent	Cumulative Percent
9	1	4	4
12	1	4	8
14	1	4	12
17	1	4	16
19	1	4	20
20	2	8	28
21	1	4	32
23	1	4	36
24	2	8	44
25	1	4	48
26	2	8	56
28	4	16	72
29	1	4	76
30	2	8	84
31	1	4	88
34	1	4	92
37	1	4	96
38	1	4	100

Posttest Scores
(N=25)

Value	Frequency	Percent	Cumulative Percent
13	1	4	4
15	1	4	8
18	1	4	12
20	1	4	16
21	1	4	20
22	1	4	24
24	1	4	28
26	1	4	32
27	1	4	36
28	1	4	40
30	1	4	44
31	1	4	48
33	3	12	60
34	2	8	68
35	2	8	76
36	1	4	80
38	2	8	88
39	1	4	92
40	2	8	100

Grouped Frequency Tables

In the example above, the tables occupy a fair amount of space because they contain a large number of different values. With an even larger number of categories, such tables would become too cumbersome. To overcome this problem, you can group the different values together into classes. There are three steps to this process:

1. Calculate the size of the range of the scores; that is, subtract the smallest value from the largest.
2. Within that range, establish between 4 and 15 classes of equal size; classes should be mutually exclusive and have a logical size.
3. Count the number of scores in each class.

In the example, scores range from 9 to 40; the size of the range is 31. A class interval of 5 would produce about 7 classes. Seven is a reasonable number of classes, and a class interval of 5 is logical and easy to handle. This produces the grouped frequency distributions in Figure 14-5.

Figure 14-5

Knowledge Test Scores Presented in Grouped Frequency Distributions

Pretest Scores		Posttest Scores	
Classes	Frequency	Classes	Frequency
6-10	1	6-10	0
11-15	2	11-15	2
16-20	4	16-20	2
21-25	5	21-25	3
26-30	9	26-30	4
31-35	2	31-35	8
36-40	2	36-40	6

There is no one correct grouped frequency distribution for such scores. For example, if you want more detail, you could use a class interval of 3 and have more classes.

Bar Graph or Histogram

Grouped frequency distributions can easily be turned into bar graphs or histograms for greater visual impact. In general, you may find it useful to create bar graphs of the most important outcomes that you wish to emphasize. (See example in Chapter 15.)

Measures of Central Tendency

Commonly you will find it useful to further summarize a set of scores by calculating a measure of central tendency or average. These averages do not present as much information as frequency distributions, but they are obviously very convenient summaries of the data. There are three different measures of central tendency.

Mode. The mode is the value with the largest frequency. In the example above, the mode for the pretest is 28; for the posttest, 33.

Note that finding the mode does not require either ordering the cases or performing any arithmetic operations on the scores. Thus, it is the only measure of central tendency that you can use with nominal data. You can also use it with ordinal, interval, and ratio variables, although it is not as good for these kinds of variables as the following measures.

Median. The median is the value of the middle score after the scores have been ordered. If there is an odd number of scores, then there is only one score in the middle and that is the median. For example, if there are 25 scores, the median is the value of the 13th score in the array. If there is an even number of scores, the median is the average of the two scores in the middle. In the example, the medians of the pretests and posttests are 26 and 33 respectively.

Finding the median does require the scores to be in order, and thus it is generally the best measure of central tendency to calculate with ordinal data. You can also use it with interval and ratio data, although in those cases the mean is normally better.

In the pretest scores above, if you decreased or increased any of the scores below 26 without letting them exceed 26, the median would not change. Similarly, if you decreased or increased any of the scores above 26 without letting them fall below 26, the median still would not change. Thus, the median is not a very sensitive measure and wastes information. Normally this insensitivity is not desirable. However, if you have interval data with one or a couple of very extreme scores that would greatly distort the mean, the insensitivity of the median would make it a better measure of central tendency.

Mean. The mean is the average of the scores and is obtained by calculating the sum of the scores and dividing by the number of scores. In the examples above, the means of the pretests and posttests are 621/25=24.84 and 743/25=29.72 respectively.

The mean does involve addition; thus, the data must be either interval or ratio. For these two, it is commonly the best measure of central tendency to calculate. Once again, if you have ordinal variables with categories that appear to be equal, then you can consider calculating the mean, although the median would be a more rigorously correct measure.

In contrast to the median, the mean is affected by all the scores, because in the first step it adds every value. Thus, it is normally a better measure of central tendency than either the mode or median. However, as noted above, if an extreme score would greatly distort it in a misleading way, the median would be better.

Using measures of central tendency in comparisons. Normally in evaluation, you will not be concerned with a single mode, median, or mean; rather, you will want to compare one mode with another, one median with another, etc. In the example above, neither the pretest nor the posttest mean is particularly useful alone. It is the comparison between the pretest and the posttest means that is useful.

Chapter 3, which discusses experimental designs, discusses proper methods of comparing different means in different experimental designs.

Measure of Dispersion

Just as describing the central tendency of data by finding the mode, median, or mean is commonly useful, so is measuring the extent to which the scores are dispersed or spread apart. For example, the mean temperature in the summer may be 80 degrees, but if the weather ranges from 60 to 110 degrees, people may be much less comfortable than if it ranges from 75 to 85 degrees. Similarly, if everyone in your class has about the same score on the test, then you can target your teaching to that particular level. On the other hand, if some students perform very well and others very poorly on a pretest, you will need to consider their varying skills.

You can obtain some estimate of the dispersion of scores by observing either the frequency distribution or the grouped frequency distribution. However, sometimes it is more convenient to summarize the dispersion with a single number.

Range. The range is simply the largest observed score minus the smallest observed score. Thus, the pretest and posttest scores have ranges of 29 and 27 respectively. In order to find the range, you must be able to order the numbers; thus, the data must be either ordinal, interval, or ratio.

Variance. To calculate the variance, find the distance of each score from the mean, square the distances, and find the mean of the squared distances. If all the scores are close together, the sum of the distances between each score and the mean will be small and the variance will be small. In contrast, if the scores are spread out considerably, the distances from the scores to the mean will be larger and the variance will be larger. Since the variance is affected by all the scores, not only by the end scores, it is a more sensitive and useful measure of dispersion than the range.

Variance involves adding and subtracting scores, so the data must be interval or ratio.

Sometimes statistical formulas or computer outputs require or provide the standard deviation. It is simply the square root of the variance. Thus, if the variance is 4, the standard deviation is 2.

Inferential Statistics

Suppose you completed a 50-item multiple-choice knowledge test without reading the questions. That is, you marked off one answer on each multiple choice question, but you had no idea what the correct answer was, because you had not read the question. If each question had 5 possible answers, you might guess correctly about 10 of them. If you repeated the test, you might guess correctly only 6 of them or 15.

This example is realistic in some ways. People do guess on tests. Sometimes they are lucky and guess many questions correctly; other times they are unlucky and guess most questions incorrectly. Moreover, during actual testing of groups, many other chance factors (e.g., respondents' having colds) affect the overall scores.

If the mean score on your posttest is better than the mean score on the pretest, you can describe the posttest mean as higher than the pretest mean. But descriptive statistics alone cannot tell you whether the improvement was probably caused by the program or by a myriad of chance factors. However, you can determine whether chance factors or the program probably produced your results by using inferential statistics, and in particular, tests of significance.

Obtaining the following tests of significance is normally easier on a computer than by hand. If you do wish to calculate them by hand, consult one of the statistics books referenced at the end of the chapter. Calculating a t-test by hand is not very difficult if you have a small number of cases.

Tests of Significance

T-tests. The t-test provides the probability that the difference between two means could have occurred by chance. Consequently, whenever you want to know whether or not the difference between two means is statistically significant, you should conduct a t-test.

T-tests require interval or ratio data. However, if the data are ordinal, have a small number of categories, and appear to have roughly equal intervals, then you can safely conduct a t-test. Technically, the t-test also requires that the distribution of each sample be normal and that the two samples have similar standard deviations. However, both of these requirements can be violated without much effect if the sample sizes are similar or are larger than 50. Because you are likely to have similar sample sizes in your analysis, you will commonly be safe using the t-test.

You can use the t-test in several different experimental designs. If you have a pretest-posttest design, use the t-test to determine whether the difference between the pretest and posttest means is significant.

- If the pretests and posttests of each case cannot be matched, use a separate or independent samples t-test.

- If cases can be matched, use the matched pairs t-test; it is both more powerful and more valid.

For a classical experimental design that includes experimental and control groups:

- If you have independent sample data and not matched pairs data, find the mean improvement in the experimental group (by subtracting the pretest mean from the posttest mean) and then compare it with the improvement in the control group.

- If you do have matched pairs data, subtract the pretest score from the posttest score for each individual and compare the mean improvement of the experimental group with that of the control group.

Analysis of variance. The analysis of variance provides the probability that the differences among two or more means could have occurred by chance. If you have only two means, the analysis of variance will give you the same probability as the t-test. In fact, the t-test is a special case of the analysis of variance. Because t-test programs are easier to run, you should use t-tests with only two means.

If you have three or more means, use the analysis of variance to determine whether all of them significantly differ from each other. It is especially well suited to comparing several different experimental groups or several different control groups (e.g., Solomon four-group design).

If you wish to observe the relative impact upon different groups of people, some of whom participated in the program and some of whom did not, use two-way analysis of variance. For example, you may have given the program to half the freshman class, half the sophomore class, half the junior class, and half the senior class. Thus, you have an experimental group for each of the four classes and want to determine which class was most affected. A discussion of two-way analysis of variance is beyond the scope of this handbook, but can be found in Blalock (1972).

Levels of Significance

When you conduct any of the tests of significance above, you will obtain a number representing the probability of the data occurring by chance. That is, the number specifies the probability that you would have obtained these scores even if the groups were not different. In the example used throughout this chapter, probability = .028. This means that even if the program had no impact, you would obtain a difference between your pretest and posttest means of at least 4.9 in 28 times out of 1,000. Because 28 times out of 1,000 is a small number of times, you can conclude that chance factors probably did not produce the difference between the pretest and posttest means and be right 972 times out of 1,000.

When you obtain your probability, you should round it upward to a level of significance. In inferential statistics, there are three important levels of significance: .05, .01, and .001.

- If the probability you obtained is greater than .05, state that the results are not statistically significant.
- If the probability you obtained is less than or equal to .05, and greater than .01, state that the results are statistically significant at the .05 level.
- If the probability you obtained is less than or equal to .01, and greater than .001, state that the results are statistically significant at the .01 level.
- If the probability you obtained is less than or equal to .001, state that the results are statistically significant at the .001 level.

The .05 level of significance means that the probability of the your data being produced by chance alone is 5 chances out of 100 or less. Conversely, it means that the chances that the program or some other important factor (other than luck) produced the change are 95 out of 100 or more. Similarly, the .01 level of significance means that chance factors would produce these results only 1 time out of 100 or less; the .001 level, 1 time out of 1,000 or less. Clearly, .001 is better than .01, and both are better than .05.

There is nothing magical about these particular levels of significance. They have simply been selected by convention.

Meaningfulness of Results

Tests of significance can tell you whether your program probably had an impact. However, they cannot tell you anything about the magnitude or importance of that impact. Particularly if your sample size is large, the program could have a very small but statistically significant impact. Thus, as a very important last step in your analysis, you should examine the magnitude of the impact and consider its importance.

If you have used criterion referenced methods, then you should give considerable weight to the percentage of people that meet the desired levels and also to the increase in the percentage of people who meet these levels.

If you chose not to specify levels of competence, then you should look at the change in the mean scores. Often it is useful to view the increase in mean scores as a percentage of the possible range. For example, if you use a 1-5 Likert scale to measure clarity of values, an increase from 3.0 to 4.0 would represent 20% of the possible range and would be substantial, whereas an increase from 3.1 to 3.2 would represent only 2% of the possible range and would be small. On a 50-item knowledge test, an increase from 33 to 34 would be small, whereas an increase from 33 to 40 would be substantial.

When making assessments about the magnitude of the change, you should be realistic. Innumerable evaluations of social programs have demonstrated that changing people's social skills and behavior is very difficult, and in general, you should be pleased with small changes. Even if you are not doing a rigorous cost/effectiveness analysis of your program, you should at least consider the cost and comprehensiveness of your program. If your program is short, you can be pleased with smaller gains; if your program is more comprehensive, you should have higher expectations.

Recommended Statistics Books

Blalock , H.M., Jr. Social Statistics. Second Edition. New York: McGraw-Hill, 1972.

Bohrnstedt, G.W., & Knoke, D. Statistics for Social Data Analysis. Itasca, Ill.: F.E. Peacock, 1982.

Freeman, L.C. Elementary Applied Statistics. New York: John Wiley and Sons, 1965.

Loether, H.J., & McTavish, D.G. Inferential Statistics for Sociologists: An Introduction. Boston: Allyn and Bacon, 1974.

Nie, N., Hull, C.H., Jenkins, J., Steinbrenner, K., & Bent, D. Statistical Package for the Social Sciences. New York: McGraw Hill, 1975.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

CHAPTER 15

WRITING THE EVALUATION REPORT

The quality of writing of your report may substantially affect the extent to which it is read and used. If you carefully describe your program, your research steps, your results, and your conclusions, your report may affect subsequent programs and policy. If you fail to present needed evidence or to focus your conclusions, your entire evaluation may have far less impact than it deserves.

Planning the Writing Project

Just as there are important steps in designing questionnaires, so there are important steps in writing a report.

- Define your audience and consider its needs and interests.
- Determine which points or conclusions you wish to emphasize, and create a detailed outline of the report.
- Write a draft, let it sit for several days or so, and then rewrite it.
- Have others involved in your research and in your field review it, and incorporate their suggestions.
- If feasible, have a professional editor help you organize and edit it.

When you consider the needs of your audience, remember that some important members of your audience will view research methodology and statistical analysis as foreign languages. Thus, you must keep in mind not only their needs for the results (to justify funding, improve the program, etc.), but also their familiarity with methods and statistics. Although you may be writing the report after having been been immersed in the fine distinctions of statistical analysis, many of your readers will be put off by them.

In some situations, you may want to summarize in lay terms the important findings and recommendations at the beginning of each section and then move into more technical explanations. Those who do not need or want the more technical material can skip over it, and those who are interested can skip to the tables, graphs, and more technical discussions of findings.

In other situations, you may find it useful to write two reports: one for general dissemination such as in your community, and another for the professional literature. You do not have to satisfy all requirements of your varied audience with one piece of writing.

As you write, try to follow these general guidelines:

- Keep both sentences and paragraphs short.
- Use descriptive section headings.
- Use active, not passive, verbs. (Not "questionnaires were completed by students"; rather, "students completed questionnaires.")

- Use definite, specific, concrete language.
- Continually edit to omit needless words, tighten loose sentences, and rewrite jargon.

Contents of the Report

Your report should explain and discuss the following:

- The background and purpose of the evaluation: who wanted it for what reasons
- The nature of the program: demographic characteristics of participants, goals, content, time period, etc.
- The specific goals or objectives that are being evaluated
- The general methods used in the evaluation
- The questionnaires used
- The sample: selection criteria, size, and other characteristics
- The results both in table form and in prose
- The limitations of the evaluation
- The conclusions and recommendations.

Presenting Quantitative Results

You should emphasize your most important points by creating one or more tables and graphs. Tables are eye-catching; graphs are even more powerful. Some people will read only the tables and graphs and ignore much of the text. Thus, tables and graphs should include your major findings and should be self explanatory.

Following are some guidelines for creating these tables:

- Create a title that accurately describes the content of the table.
- Provide pretest and posttest means to show the size of the increase or change.
- Include the sample size.
- Include tests of significance if they were calculated. Follow convention by letting * = a result significant at the .05 level; ** = a result significant at the .01 level; and *** = a result significant at the .001 level.
- Rather than using numbers to signify footnotes, use letters (e.g., a, b, c).

For example, the data from the previous chapter has been presented in Table 15-1.

If you have more than one group, then you need to add additional rows for those groups. For example, Table 15-2 is based upon different fictitious data with additional groups.

Bar Graphs

Since bar graphs have a greater visual impact than tables, you may find it useful to create bar graphs of the most important outcomes. Grouped frequency distributions and the means and medians of different groups lend themselves well to this treatment.

Table 15-1

Mean Pretest and Posttest Scores on a 40-Item Multiple Choice Test

Group	Sample Size[a]	Pretest	Posttest	Difference Between Means	T-test for Difference Between Means
Sexuality Education Class	25	24.84	29.72	4.88	2.27**

**Significant at the .01 level.

[a]Matched pairs.

Table 15-2

Mean Pretest and Posttest Scores on a 40-Item Multiple Choice Test for a Sexuality Education Class and Its Control Group

Group	Sample Size	Pretest	Posttest	Difference Between Means	T-test for Difference Between Means
Sexuality Education Class	43	29.45	33.93	4.48	3.18**
Control Group	42	28.22	31.21	2.99	

**Significant at the .01 level.

Following are some guidelines for creating these tables:

- Create a title that accurately describes the content of the table.
- Provide pretest and posttest means to show the size of the increase or change.
- Include the sample size.
- Include tests of significance if they were calculated. Follow convention by letting * = a result significant at the .05 level; ** = a result significant at the .01 level; and *** = a result significant at the .001 level.
- Rather than using numbers to signify footnotes, use letters (e.g., a, b, c).

For example, the data from the previous chapter has been presented in Table 15-1.

If you have more than one group, then you need to add additional rows for those groups. For example, Table 15-2 is based upon different fictitious data with additional groups.

Bar Graphs

Since bar graphs have a greater visual impact than tables, you may find it useful to create bar graphs of the most important outcomes. Grouped frequency distributions and the means and medians of different groups lend themselves well to this treatment.

Following are several suggestions for creating bar graphs:

- Use a separate bar (or pair of bars) for each major objective.
- Have bars which should be compared next to one another. For example, place the bar representing a pretest next to the bar representing the posttest.
- Label each axis.
- Make intervals along the axis equal.
- If possible, use a zero point at the base of the axes; if not, put a jagged line through both the axis and the bars near the base.
- Above each bar, indicate the actual number the bar represents (e.g., frequencies or means).
- Represent the criterion or goal of an objective by drawing a line that is perpendicular to the bars (see Figure 15-2).

The frequency distributions and the means for the data from the previous chapter can be represented by bar graphs as in Figure 15-1 and Figure 15-2 respectively.

Presenting Nonquantitative Data

In the analysis of sexuality education programs, the most common form of non-quantitative data is written statements. These may be unsolicited or the result of open-ended questions (What did you learn in this program? How did this program affect you? What changes in the program would you recommend?). Too often such statements are neither fully analyzed nor reported, despite their potential value.

Figure 15-1

Number of People Receiving Different Scores on Pretests and Posttests

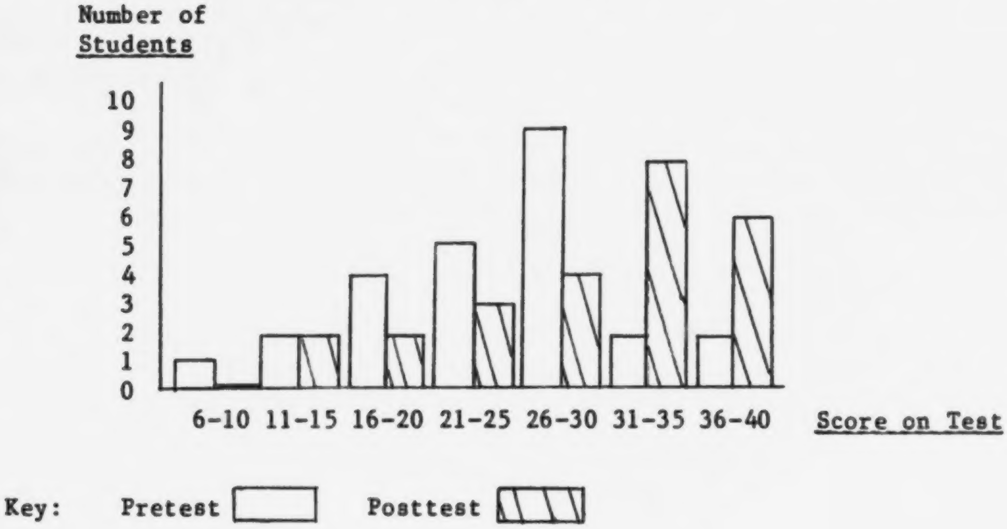
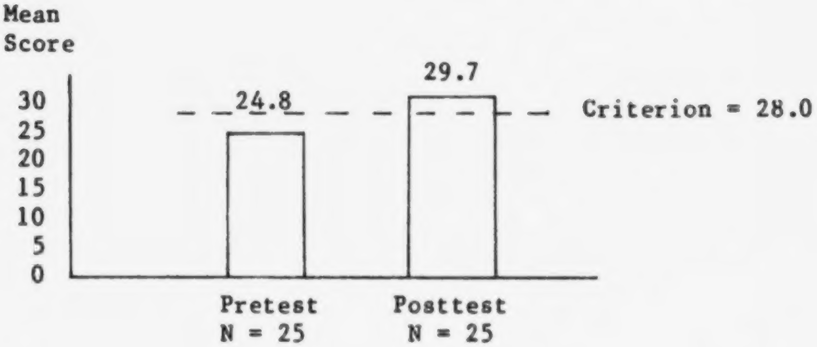


Figure 15-2

The Mean Knowledge Test Scores for Students on the Pretest and Posttest



When reporting such statements, it is extremely important to fairly represent negative as well as positive statements. To report only positive statements and to exclude negative statements is simply misleading and invalid. Just as it is unethical to throw out low scores from a posttest knowledge tests, so it is also unethical to fail to proportionately report negative statements. Moreover, critical statements can help others recognize deficiencies in programs and then improve them.

An assortment of spontaneous or solicited questions can appear to defy reasonable organization. If so, begin by writing each statement on a separate sheet of paper (e.g. 3 x 5 cards); then organize the statements by content. You may then do one or more of the following:

- Select a few representative statements to illustrate points made in the quantitative analysis or in the text.
- Include all the statements in the report.
- Accurately summarize the statements by presenting the frequency with which different themes or ideas are mentioned.

Dilemmas in Writing and Publishing the Results

Evaluators often do not obtain the results they hoped for. They are then faced with the dilemma of what to report. Should they report only the positive findings, or should they report all findings accurately? Especially when the findings have political significance, there can be considerable pressure to emphasize only the positive findings, and people will often offer superficially convincing justifications for doing so. However, in the long run, burying negative findings is not helpful to either sexuality education or evaluation research. You certainly have an ethical obligation to report findings accurately. Moreover, if you find negative results, you should report them so that people can try to improve programs and so that other findings can be trusted.

This responsibility does not necessarily mean that you should not give any thought to the possible political use of your findings. People both for and against sexuality education may selectively quote your results, and as much as possible, you should not write statements that will be greatly misleading if quoted out of context. In other words, each statement should be balanced as much as possible.

The same considerations arise when you consider publishing your results. If those people who find that sexuality education improves behavior publish their results, and if those people who find that sexuality education has no impact or a negative impact fail to publish their findings, then the literature will obviously become biased and people will incorrectly believe that sexuality education is more effective than it actually is.

On the other hand, if you failed to obtain positive findings and have good reason to believe that your findings are invalid, than you should either reevaluate your program before publishing the results, or you should emphasize in your report that your results may be invalid.

Suggested Readings

Bernstein, T.M. *The Careful Writer: A Modern Guide to English Usage*. New York: Atheneum, 1977.

A Manual of Style, 12th ed., rev. Chicago: University of Chicago Press, 1969.

Morris, L.L., & Fitz-Gibbon, C.T. *How to Present an Evaluation Report*. Beverly Hills, Calif.: Sage Publications, 1978.

Publication Manual of the American Psychological Association, 2d ed. Washington, D.C.: American Psychological Association, 1974.

Strunk, W., Jr., & White, E.B. *Elements of Style*, 2d ed. New York: MacMillan, 1972.

CHAPTER 16

EVALUATING SPECIFIC KINDS OF PROGRAMS

Although the methods described in this volume are generally applicable to many kinds of sexuality education programs, some programs have special characteristics that affect their evaluation. This chapter discusses some of the special considerations for evaluating specific kinds of programs.

Comprehensive Programs Lasting About a Semester

The methods already described in this volume are well suited to evaluating comprehensive programs. Because most comprehensive programs last numerous weeks, using an experimental design with a control group is especially important; during the elapsed time, students' knowledge, attitudes, and behavior may change even if they do not take the program. Comprehensive programs are also more likely than shorter programs to have a variety of effects upon knowledge, attitudes, and behavior. Thus, it is especially important that you carefully specify objectives and measure most or all of the possible outcomes.

At the end of a semester course, students may trust the teacher much more and may be much more open about their sexuality than at the beginning of the semester. Thus, they may answer questions about attitudes and behavior more honestly at the end of the semester than at the beginning. To reduce this possible bias, you can use the first week of the course to increase trust and openness, but not teach much about sexuality, and then administer the pretests during the second week.

Short Structured Courses Lasting 1 or 2 Weeks

Programs that last a relatively short time are likely to have less impact than programs lasting a semester or longer. Thus, there are fewer plausible outcomes that you should measure and the questionnaires should be shorter. For example, short courses are less likely to have an impact upon self esteem and there is less need to measure self esteem.

Programs that last 1 or 2 weeks are unlikely to produce much behavioral change during the course but may produce considerable behavioral change after the course. Thus, you should be sure to administer second posttests weeks or months after the end of the course.

One-day Conferences

Because 1-day conferences are so short, they require a modified experimental design. Many desired behavioral outcomes of the conference cannot take place during that day and need not be measured at the end of the conference. Obviously, for example, there is no need to measure the amount of unprotected sexual activity twice on the day of the conference. Thus, posttests measuring behavior should be administered weeks or months later. On the other hand, both knowledge and attitudes may change during the day and can profitably be measured at both the beginning and the end of the conference.

The short duration of the conference has a second impact upon the methodology. At the beginning of the day, participants are usually fresh and relatively willing to complete questionnaires carefully. By the end of the day, they are likely to be tired, to have less energy and concentration, and to be less willing to complete lengthy questionnaires carefully. Thus, if you do administer questionnaires at the end of the day, you should make the questionnaires as short and easy as possible.

Because the short posttests administered at the end of the conference are not sufficient to measure change in many important outcomes, administering questionnaires at a later time becomes especially important. Doing this usually requires obtaining the participants' names and addresses and making some arrangements for mailing them the questionnaires. Especially when the questionnaires contain sensitive questions about sexuality, follow these procedures:

- Carefully explain to the participants the importance of their completing the second posttests at home.

- Obtain their permission to send the questionnaires to their homes.

- Strongly encourage them to complete the questionnaires anonymously without anyone else's help or advice.

- Encourage participants to return the questionnaires, even if they decide not to complete them, so that questionnaires with sensitive questions do not circulate throughout the community.

Many conferences are voluntary, and some participants may leave the conference before it ends. Participants' leaving before completing the posttest questionnaires may affect your results negatively. First, it will reduce your sample size. Second, the participants who leave may be different in some way (in intelligence, motivation, or satisfaction with the conference) from those who stay. Thus, their loss may bias your sample and affect its representativeness. This possible bias might render a comparison of the mean scores on the pretest and posttest invalid. To prevent such a bias, use identification numbers and match pretests with posttests.

Similar biases may occur when you obtain delayed posttest data; less motivated people may refrain from returning questionnaires. Once again, you can reduce this problem by using identification numbers and matching pretests with the delayed posttests.

Because 1-day conferences are short, they probably will have less impact than longer, more comprehensive programs. Thus, you need not measure many of the outcomes that you might wish to measure in longer programs. For example, you probably need not measure changes in self esteem or skills, because conferences are not likely to have an impact upon these outcomes.

Some conferences have flexible, unstructured formats. That is, participants can attend different activities, peruse materials on their own, ask counselors or other professionals questions during small group discussions, etc. Thus, participants may be less likely to learn a specific set of facts, but more likely to learn particular factual information that is immediately relevant to them. Thus, traditional knowledge tests may be a poor method of assessing knowledge gain. Unfortunately, there are not many good alternatives. You can include open-ended questions that ask participants 1) to summarize what they have learned or 2) to specify several factual pieces of information that they learned, but such questions usually fail to accurately assess how much the respondents learned, how much they actually know, and what topics should be covered more fully.

Peer Education Programs

Peer education programs are particularly difficult to evaluate because their effects are likely to be small and diffuse. That is, peer educators are likely to interact with only a small number of students who are scattered throughout the student body. Nevertheless, there are at least two potentially successful strategies for evaluation.

First, you can collect data on the entire school body during the years both before and after the peer education program is implemented. This data can be questionnaire data collected from the students, pregnancy data collected from clinics and doctors, or other kinds of data. A comparison of the before and after data will give some evidence for the effects of the program.

Second, if the peer educators speak before selected classes of students in the school, and if they subsequently meet primarily with students in these classes, then you can randomly assign classes to experimental and control groups. The classes to which the peer educators speak would comprise the experimental group; the classes to which they don't speak, the control group. You can then administer questionnaires to both groups of classes at the beginning and end of the year and compare the changes over time.

Questionnaires should include questions on the number of contacts that the students had with the peer educators. If some students had no contact with the peer educators, then changes in their knowledge or attitudes could not have been produced by interaction with the peer educators. You should, however, use this line of reasoning with caution. Do not compare students who seek information from the peer educators with students who do not, because students who meet with the peer educators may be more likely to be sexually active and in need of information than those who do not seek information.

Your questionnaires should be quite short because the small amount of interaction with the peer educators is not likely to produce a substantial amount of change and you do not need to measure as many outcomes.

In general, all analyses of peer education programs should be viewed with caution:

- The effects of interaction with peer educators are small and diffuse.

- Assessing the amount of interaction between each student and the peer educators is difficult.

- Students who seek advice from peer educators are likely to be different from those who don't.

Parent/Child Programs

Programs offered to parents and their children together can be evaluated in much the same way that other programs have been evaluated. However, such programs offer an additional possibility of matching the parents with their children, asking similar questions of each, and then comparing answers. For example, to measure the impact of the program on family communication about sexuality, you can ask both parents and their children how often they talk about sexuality and how comfortable they are during those conversations. (To date, however, researchers have found very little relationship between the reports of the parents and the reports of their children.)

A few researchers have tried to measure the impact of parent courses upon the quality of family communication by video or audio taping family conversation about sexuality both before and after the program. Although intriguing, we don't recommend the method.

- The presence of the tape recorder and the instructions for the session prevent people from talking in a normal manner. People find taping far more intimidating than completing questionnaires.
- For some families, the instructions to talk about sexuality encourages or forces them to do something they have never done before and may therefore give an inaccurate picture of family communication.
- Many families do not even talk about sexuality during their preprogram tape recordings, rendering those tapes invalid.
- Implementing the recording sessions is time consuming to both the participants and the researchers because the recordings cannot be completed simultaneously in a single group.
- Coding tape recordings is very difficult.

Conclusions

We have been evaluating sexuality education programs for several years and have learned a great deal from our experiences. At the beginning we certainly made our share of mistakes: we designed questionnaires that were too difficult and too long; we included a few questions that were too sensitive for some people; we tried to measure too many outcomes; we sometimes failed to obtain data from control groups; we sometimes failed to ensure the proper administration of questionnaires; we tried a few new, innovative, and totally unworkable approaches. However, we learned, continually improved our methods, and collected very useful data on the effects of programs.

We have written this volume so that you can learn from some of our experiences and avoid some of our mistakes. No single volume can present all that you need to know to conduct valid evaluations. However, if you follow the principles and methods described in this volume you are likely to obtain valid and very useful information about the success of your program. You can then improve or expand your program.

Remember the following points:

- Specify clearly the most important outcomes of the program that you wish to measure. Recognize that you probably cannot measure validly all important outcomes; be realistic about what you can measure. Be careful about asking sensitive questions.
- Obtain approval from the school or organizational authorities, parents, participants, and other appropriate groups. Be able to justify why you're asking each question.
- Use multiple methods as much as possible; design different kinds of questionnaires, give them to different groups of people, and use other methods as well.
- Use as many characteristics of experimental designs as possible. At a minimum collect pretest and posttest data.
- Repeatedly pretest your questionnaires and be sure they are reliable and valid for your particular respondents.
- Make sure the administration of the questionnaires is rigorous; if students fail to treat the questionnaires seriously, the most carefully designed evaluation can be useless.
- Be especially careful to maintain the anonymity or confidentiality of any potentially sensitive data.
- Be prepared to learn that your evaluation indicates your program is not as effective as you had hoped.
- At every stage guard against letting your hopes and values bias your evaluation.

Increasing numbers of people are evaluating their programs with these methods, improving their programs, and finding that the combination of evaluation and program improvement is well worth the effort.

If you have conducted few or no evaluations before, you may feel intimidated by all the methods described in this volume. If so, remember some of the suggestions given for making the process easier:

- Start with a small and relatively simple evaluation, and as you become more familiar with evaluation methods, improve the size and quality of your evaluation.
- Contact methodological consultants or other members in the field who have previously completed evaluations, ask them questions, and learn from their experiences.
- At first use questionnaires and other materials developed by others; later develop your own questionnaires.

Remember, both evaluation methods and statistics may appear difficult, but they are a logical set of procedures to collect data systematically and make that data more clear and concise so that important questions can be better answered. In many respects learning to use these methods is like learning to ride a bicycle; the going is difficult at first, but becomes much easier with practice. Just as falling down makes the principles of bike riding more obvious, so will making mistakes with these methods make their rationale more obvious.

APPENDIX

QUESTIONNAIRES

These questionnaires are modified versions of the questionnaires that we have used to evaluate sexuality education programs for adolescents. They include evaluations and assessments of the course to administer at the end of the course and questionnaires which measure knowledge, attitude, and behavior to administer before and after the course. See the guidelines below.

Questionnaire	Administration
Knowledge Questionnaire	Before and after the course
Attitude and Value Inventory	Before and after the course
Behavior Inventory	Before and after the course
Knowledge, Attitude, and Behavior Inventory (An integrated, condensed version of the first three)	Before and after the course
Course Evaluation	After the course
Assessment of Course Impact	After the course
Course Assessment for Parents	After the course

These questionnaires provide examples of questions that you can use. You should modify the questionnaires to meet the values of your community, the particular goals of your program, and the characteristics of your program participants. For example, if the adolescents in your program are not likely to be sexually active, then you should remove those questions dealing with sexual activity. You should also consider the appropriate length of each questionnaire. If the questionnaire is too long, remove questions or use the Knowledge, Attitude, and Behavior Inventory which contains the most important questions from the Knowledge Questionnaire, the Attitude and Value Inventory, and the Behavior Inventory.

The Knowledge Questionnaire, Attitude and Value Inventory, and Behavior Inventory can be subdivided into the following individual scales. The questions on the Knowledge Questionnaire and the Behavior Inventory can be analyzed separately or as scales. Although the scales include questions measuring the same topics, they are not true multi-item scales. In contrast, the scales in the Attitude and Value Inventory are true multi-item scales. Thus, if you intend to measure a particular attitude, you should include all the questions of that scale. That is, you should not use the questions individually.

For reasons of space, we have listed here only the item numbers for each scale. Because the Attitude and Value Inventory contains true multi-item scales, you may prefer to read the items grouped as scales rather than randomly ordered through the inventory. For your convenience in doing so, we have grouped all the items by scale at the end of that questionnaire.

Scales in the Knowledge Questionnaire	Question Numbers
Physical Development	2, 8, 13, 15, 25, 28
Adolescent Relationships	22, 27, 29
Adolescent Sexual Activity	1, 3, 16, 17
Adolescent Pregnancy	6, 20, 23
Adolescent Marriage	9, 30
Probability of Pregnancy	5, 10, 12, 19
Birth Control	4, 11, 18, 26, 31, 32, 34
Sexually Transmitted Disease	7, 14, 21, 24, 33

Scales in the Attitude and Value Inventory	Question Numbers
Clarity of Long Term Goals	10, 23, 30, 37, 51
Clarity of Personal Sexual Values	5, 13, 25, 49, 70
Understanding of Emotional Needs	14, 17, 48, 56, 62
Understanding of Personal Social Behavior	6, 19, 27, 34, 66
Understanding of Personal Sexual Response	21, 31, 36, 45, 52
Attitude Toward Various Gender Role Behaviors	8, 28, 41, 50, 65
Attitude Toward Sexuality In Life	12, 42, 55, 58, 64
Attitude Toward the Importance of Birth Control	4, 16, 40, 59, 61
Attitude Toward Premarital Intercourse	2, 20, 22, 29, 63
Attitude Toward the Use of Pressure and Force in Sexual Activity	9, 15, 46, 47, 54
Recognition of the Importance of the Family	11, 24, 53, 60, 69
Self Esteem	3, 26, 35, 44, 68
Satisfaction with Personal Sexuality	7, 18, 33, 39, 57
Satisfaction with Social Relationships	1, 32, 38, 43, 67

Scales in the Behavior Inventory	Question Numbers
Taking Responsibility for Behavior	1, 2
Decisionmaking Skills	3, 4, 5, 6
Decisionmaking Skills about Sexual Behavior	7, 8, 9, 10, 11
Communication Skills	12, 13, 14, 15, 16, 17, 18, 19
Assertiveness Skills about Sexual Behavior	20, 21, 22, 23, 24
Comfort with Social Interaction	25, 26, 27, 28
Comfort Talking about Sex and Birth Control	29, 30, 31, 32, 33, 34, 36
Comfort Talking about Sexuality with Parents	31, 34
Comfort Talking about Sexuality with Friends	29, 32
Comfort Talking about Sexuality with Girl or Boyfriend	30, 33
Comfort Expressing Concern and Caring	35
Comfort Being Assertive Sexually	36, 37
Comfort with Current Sex Life	38
Comfort Getting and Using Birth Control	39, 40, 41, 42
Sexual Activity	43, 44, 45
Use of Birth Control	46, 47, 48
Frequency of Communication about Sex and Birth Control with Parents	49, 52
Frequency of Communication about Sex and Birth Control with Friends	50, 53
Frequency of Communication about Sex and Birth Control with Boyfriend or Girlfriend	51, 54

KNOWLEDGE QUESTIONNAIRE

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ________________________

Teacher's name: ________________________________

Your birth date: Month _________ Day ______

Your sex (Check one): Male ___ Female ___

Your grade level in school (Check one): 9 _____
10 _____
11 _____
12 _____

Please circle the one best answer to each of the questions below.

1. By the time teenagers graduate from high schools in the United States:

   a. only a few have had sex (sexual intercourse).
   b. about half have had sex.
   c. about 80% have had sex.

2. During their menstrual periods, girls:

   a. are too weak to participate in sports or exercise.
   b. have a normal, monthly release of blood from the uterus.
   c. cannot possibly become pregnant.
   d. should not shower or bathe.
   e. all of the above.

3. It is harmful for a woman to have sex (sexual intercourse) when she:

   a. is pregnant.
   b. is menstruating.
   c. has a cold.
   d. has a sexual partner with syphilis.
   e. none of the above.

4. Some contraceptives:

   a. can be obtained only with a doctor's prescription.
   b. are available at family planning clinics.
   c. can be bought over the counter at drug stores.
   d. can be obtained by people under 18 without their parents' permission.
   e. all of the above.

5. If 10 couples have sexual intercourse regularly without using any kind of birth control, the number of couples who become pregnant by the end of 1 year is about:

   a. one.
   b. three.
   c. six.
   d. nine.
   e. none of the above.

6. When unmarried teenage girls learn they are pregnant, the largest group of them decide:

   a. to have an abortion.
   b. to put the child up for adoption.
   c. to raise the child at home.
   d. to marry and raise the child with the husband.
   e. none of the above.

7. People having sexual intercourse can best prevent getting a sexually transmitted disease (VD or STD) by using:

   a. condoms (rubbers).
   b. contraceptive foam.
   c. the pill.
   d. withdrawal (pulling out).

8. When boys go through puberty:

   a. they lose their "baby fat" and become slimmer.
   b. their penises become larger.
   c. they produce sperm.
   d. their voices become lower.
   e. all of the above.

9. Married teenagers:

   a. have the same social lives as their unmarried friends.
   b. avoid pressure from friends and family.
   c. still fit in easily with their old friends.
   d. usually support themselves without help from their parents.
   e. none of the above.

10. If a couple has sexual intercourse and uses no birth control, the woman might get pregnant:

   a. any time during the month.
   b. only 1 week before menstruation begins.
   c. only during menstruation.
   d. only 1 week after menstruation begins.
   e. only 2 weeks after menstruation begins.

11. The method of birth control which is least effective is:

   a. a condom with foam.
   b. the diaphragm with spermicidal jelly.
   c. withdrawal (pulling out).
   d. the pill.
   e. abstinence (not having intercourse).

12. It is possible for a woman to become pregnant:

   a. the first time she has sex (sexual intercourse).
   b. if she has sexual intercourse during her menstrual period.
   c. if she has sexual intercourse standing up.
   d. if sperm get near the opening of the vagina, even though the man's penis does not enter her body.
   e. all of the above.

13. Physically:

   a. girls usually mature earlier than boys.
   b. most boys mature earlier than most girls.
   c. all boys and girls are fully mature by age 16.
   d. all boys and girls are fully mature by age 18.

14. It is impossible now to cure:

   a. syphilis.
   b. gonorrhea.
   c. herpes virus #2.
   d. vaginitis.
   e. all of the above.

15. When men and women are physically mature:

   a. each female ovary releases two eggs each month.
   b. each female ovary releases millions of eggs each month.
   c. male testes produce one sperm for each ejaculation (climax).
   d. male testes produce millions of sperm for each ejaculation (climax).
   e. none of the above.

16. Teenagers who choose to have sexual intercourse may possibly:

   a. have to deal with a pregnancy.
   b. feel guilty.
   c. become more close to their sexual partners.
   d. become less close to their sexual partners.
   e. all of the above.

17. As they enter puberty, teenagers become more interested in sexual activities because:

   a. their sex hormones are changing.
   b. the media (TV, movies, magazines, records) push sex for teenagers.
   c. some of their friends have sex and expect them to have sex also.
   d. all of the above.

18. To use a condom the correct way, a person must:

   a. leave some space at the tip for the guy´s fluid.
   b. use a new one every time sexual intercourse occurs.
   c. hold it on the penis while pulling out of the vagina.
   d. all of the above.

19. The proportion of American girls who become pregnant before turning 20 is:

   a. 1 out of 3.
   b. 1 out of 11.
   c. 1 out of 43.
   d. 1 out of 90.

20. In general, children born to young teenage parents:

   a. have few problems because their parents are emotionally mature.
   b. have a greater chance of being abused by their parents.
   c. have normal birth weight.
   d. have a greater chance of being healthy.
   e. none of the above.

21. Treatment for venereal disease is best if:

  a. both partners are treated at the same time.
  b. only the partner with the symptoms sees a doctor.
  c. the person takes the medicine only until the symptoms disappear.
  d. the partners continue having sex (sexual intercourse).
  e. all of the above.

22. Most teenagers:

  a. have crushes or infatuations that last a short time.
  b. feel shy or awkward when first dating.
  c. feel jealous sometimes.
  d. worry a lot about their looks.
  e. all of the above.

23. Most unmarried girls who have children while still in high school:

  a. depend upon their parents for support.
  b. finish high school and graduate with their class.
  c. never have to be on public welfare.
  d. have the same social lives as their peers.
  e. all of the above.

24. Syphilis:

  a. is one of the most dangerous of the venereal diseases.
  b. is known to cause blindness, insanity, and death if untreated.
  c. is first detected as a chancre sore on the genitals.
  d. all of the above.

25. For a boy, nocturnal emissions (wet dreams) means he:

  a. has a sexual illness.
  b. is fully mature physically.
  c. is experiencing a normal part of growing up.
  d. is different from most other boys.

26. If people have sexual intercourse, the advantage of using condoms is that they:

  a. help prevent getting or giving VD.
  b. can be bought in drug stores by either sex.
  c. do not have dangerous side effects.
  d. do not require a prescription.
  e. all of the above.

27. If two people want to have a close relationship, it is important that they:

  a. trust each other and are honest and open with each other.
  b. date other people.
  c. always think of the other person first.
  d. always think of their own needs first.
  e. all of the above.

28. The physical changes of puberty:

a. happen in a week or two.
b. happen to different teenagers at different ages.
c. happen quickly for girls and slowly for boys.
d. happen quickly for boys and slowly for girls.

29. For most teenagers, their emotions (feelings):

a. are pretty stable.
b. seem to change frequently.
c. don't concern them very much.
d. are easy to put into words.
e. are ruled by their thinking.

30. Teenagers who marry, compared to those who do not:

a. are equally likely to finish high school.
b. are equally likely to have children.
c. are equally likely to get divorced.
d. are equally likely to have successful work careers.
e. none of the above.

31. The rhythm method (natural family planning):

a. means couples cannot have intercourse during certain days of the woman's menstrual cycle.
b. requires the woman to keep a record of when she has her period.
c. is effective less than 80% of the time.
d. is recommended by the Catholic church.
e. all of the above.

32. The pill:

a. can be used by any woman.
b. is a good birth control method for women who smoke.
c. usually makes menstrual cramping worse.
d. must be taken for 21 or 28 days in order to be effective.
e. all of the above.

33. Gonorrhea:

a. is 10 times more common than syphilis.
b. is a disease that can be passed from mothers to their children during birth.
c. makes many men and women sterile (unable to have babies).
d. is often difficult to detect in women.
e. all of the above.

34. People choosing a birth control method:

a. should think only about the cost of the method.
b. should choose whatever method their friends are using.
c. should learn about all the methods before choosing the one that's best for them.
d. should get the method that's easiest to get.
e. all of the above.

Answers to the Knowledge Questionnaire

Question	Answer	Question	Answer
1	b	18	d
2	b	19	a
3	d	20	b
4	e	21	a
5	d	22	e
6	a	23	a
7	a	24	d
8	e	25	c
9	e	26	e
10	a	27	a
11	c	28	b
12	e	29	b
13	a	30	e
14	c	31	e
15	d	32	d
16	e	33	e
17	d	34	c

ATTITUDE AND VALUE INVENTORY

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ______________________

Teacher's name: ______________________________

Your birth date: Month _________ Day ______

Your sex (Check one): Male ___ Female ___

Your grade level in school (Check one): 9 _____
10 _____
11 _____
12 _____

The questions below are not a test of how much you know. We are interested in what you believe about some important issues. Please rate each statement according to how much you agree or disagree with it. Everyone will have different answers. Your answer is correct if it describes you very well.

Circle: 1 = if you Strongly Disagree with the statement.
2 = if you Somewhat Disagree with the statement.
3 = if you feel Neutral about the statement.
4 = if you Somewhat Agree with the statement.
5 = if you Strongly Agree with the statement.

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1.	I am very happy with my friendships.	1	2	3	4	5
2.	Unmarried people should not have sex (sexual intercourse).	1	2	3	4	5
3.	Overall, I am satisfied with myself.	1	2	3	4	5
4.	Two people having sex should use some form of birth control if they aren't ready for a child.	1	2	3	4	5
5.	I'm confused about my personal sexual values and beliefs.	1	2	3	4	5
6.	I often find myself acting in ways I don't understand.	1	2	3	4	5
7.	I am not happy with my sex life.	1	2	3	4	5
8.	Men should not hold jobs traditionally held by women.	1	2	3	4	5
9.	People should never take "no" for an answer when they want to have sex.	1	2	3	4	5
10.	I don't know what I want out of life.	1	2	3	4	5
11.	Families do very little for their children.	1	2	3	4	5
12.	Sexual relationships create more problems than they're worth.	1	2	3	4	5
13.	I'm confused about what I should and should not do sexually.	1	2	3	4	5
14.	I know what I want and need emotionally.	1	2	3	4	5
15.	No one should pressure another person into sexual activity.	1	2	3	4	5

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
16.	Birth control is not very important.	1	2	3	4	5
17.	I know what I need to be happy.	1	2	3	4	5
18.	I am not satisfied with my sexual behavior (sex life).	1	2	3	4	5
19.	I usually understand the way I act.	1	2	3	4	5
20.	People should not have sex before marriage.	1	2	3	4	5
21.	I do not know much about my own physical and emotional sexual response.	1	2	3	4	5
22.	It is all right for two people to have sex before marriage if they are in love.	1	2	3	4	5
23.	I have a good idea of where I´m headed in the future.	1	2	3	4	5
24.	Family relationships are not important.	1	2	3	4	5
25.	I have trouble knowing what my beliefs and values are about my personal sexual behavior.	1	2	3	4	5
26.	I feel I do not have much to be proud of.	1	2	3	4	5
27.	I understand how I behave around others.	1	2	3	4	5
28.	Women should behave differently from men most of the time.	1	2	3	4	5
29.	People should have sex only if they are married.	1	2	3	4	5
30.	I know what I want out of life.	1	2	3	4	5
31.	I have a good understanding of my own sexual feelings and reactions.	1	2	3	4	5
32.	I don´t have enough friends.	1	2	3	4	5
33.	I´m happy with my sexual behavior now.	1	2	3	4	5
34.	I don´t understand why I behave with my friends as I do.	1	2	3	4	5
35.	At times I think I´m no good at all.	1	2	3	4	5

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
36.	I know how I react in different sexual situations.	1	2	3	4	5
37.	I have a clear picture of what I'd like to be doing in the future.	1	2	3	4	5
38.	My friendships are not as good as I would like them to be.	1	2	3	4	5
39.	Sexually, I feel like a failure.	1	2	3	4	5
40.	More people should be aware of the importance of birth control.	1	2	3	4	5
41.	At work and at home, women should not have to behave differently from men, when they are equally capable.	1	2	3	4	5
42.	Sexual relationships make life too difficult.	1	2	3	4	5
43.	I wish my friendships were better.	1	2	3	4	5
44.	I feel that I have many good personal qualities.	1	2	3	4	5
45.	I am confused about my reactions in sexual situations.	1	2	3	4	5
46.	It is all right to pressure someone into sexual activity.	1	2	3	4	5
47.	People should not pressure others to have sex with them.	1	2	3	4	5
48.	Most of the time my emotional feelings are clear to me.	1	2	3	4	5
49.	I have my own set of rules to guide my sexual behavior (sex life).	1	2	3	4	5
50.	Women and men should be able to have the same jobs, when they are equally capable.	1	2	3	4	5
51.	I don't know what my long-range goals are.	1	2	3	4	5
52.	When I'm in a sexual situation, I get confused about my feelings.	1	2	3	4	5
53.	Families are very important.	1	2	3	4	5

		Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
54.	It is all right to demand sex from a girlfriend or boyfriend.	1	2	3	4	5
55.	A sexual relationship is one of the best things a person can have.	1	2	3	4	5
56.	Most of the time I have a clear understanding of my feelings and emotions.	1	2	3	4	5
57.	I am very satisfied with my sexual activities just the way they are.	1	2	3	4	5
58.	Sexual relationships only bring trouble to people.	1	2	3	4	5
59.	Birth control is not as important as some people say.	1	2	3	4	5
60.	Family relationships cause more trouble than they´re worth.	1	2	3	4	5
61.	If two people have sex and aren´t ready to have a child, it is very important that they use birth control.	1	2	3	4	5
62.	I´m confused about what I need emotionally.	1	2	3	4	5
63.	It is all right for two people to have sex before marriage.	1	2	3	4	5
64.	Sexual relationships provide an important and fulfilling part of life.	1	2	3	4	5
65.	People should not be expected to behave in certain ways just because they are male or female.	1	2	3	4	5
66.	Most of the time I know why I behave the way I do.	1	2	3	4	5
67.	I feel good having as many friends as I have.	1	2	3	4	5
68.	I wish I had more respect for myself.	1	2	3	4	5
69.	Family relationships can be very valuable.	1	2	3	4	5
70.	I know for sure what is right and wrong sexually for me.	1	2	3	4	5

SCALES IN THE ATTITUDE AND VALUE INVENTORY

Clarity of Long Term Goals

10. I don't know what I want out of life.
23. I have a good idea of where I'm headed in the future.
30. I know what I want out of life.
37. I have a clear picture of what I'd like to be doing in the future.
51. I know what my long range goals are.

Clarity of Personal Sexual Values

5. I'm confused about my personal sexual values and beliefs.
13. I'm confused about what I should and should not do sexually.
25. I have trouble knowing what my beliefs and values are about my personal sexual behavior.
49. I have my own set of rules to guide my sexual behavior (sex life).
70. I know for sure what is right and wrong sexually for me.

Understanding of Emotional Needs

14. I know what I want and need emotionally.
17. I know what I need to be happy.
48. Most of the time my emotional feelings are clear to me.
56. Most of the time I have a clear understanding of my feelings and emotions.
62. I'm confused about what I need emotionally.

Understanding of Personal Social Behavior

6. I often find myself acting in ways I don't understand.
19. I usually understand the way I act.
27. I understand how I behave around others.
34. I don't understand why I behave with my friends as I do.
66. Most of the time I know why I behave the way I do.

Understanding of Personal Sexual Response

21. I do not know much about my own physical and emotional sexual response.
31. I have a good understanding of my own sexual feelings and reactions.
36. I know how I react in different sexual situations.
45. I am confused about my reactions in sexual situations.
52. When I'm in a sexual situation, I get confused about my feelings.

Attitude Toward Various Gender Role Behaviors

8. Men should not hold jobs traditionally held by women.
28. Women should behave differently from men most of the time.
41. At work and at home, women should not have to behave differently than men, when they are equally capable.

50. Women and men should be able to have the same jobs, when they are equally capable.
65. People should not be expected to behave in certain ways just because they are male or female.

Attitude Toward Sexuality in Life

12. Sexual relationships create more problems than they're worth.
42. Sexual relationships make life too difficult.
55. A sexual relationship is one of the best things a person can have.
58. Sexual relationships only bring trouble to people.
64. Sexual relationships provide an important and fulfilling part of life.

Attitude Toward the Importance of Birth Control

4. Two people having sex should use some form of birth control, if they aren't ready for a child.
16. Birth control is not very important.
40. More people should be aware of the importance of birth control.
59. Birth control is not as important as some people say.
61. If two people hve sex and aren't ready to have a child, it is very important that they use birth control.

Attitude Toward Premarital Intercourse

2. Unmarried people should not have sex.
20. People should not have sex before marriage.
22. It is all right for two people to have sex before marriage if they are in love.
29. People should have sex only if they are married.
63. It is all right for two people to have sex before marriage.

Attitude Toward the Use of Pressure and Force in Sexual Activity

9. People should never take "no" for an answer when they want to have sex.
15. No one should pressure another person into sexual activity.
46. It is all right to pressure someone into sexual activity.
47. People should not pressure others to have sex with them.
54. It is all right to demand sex from a girlfriend or boyfriend.

Recognition of the Importance of the Family

11. Families do very little for their children.
24. Family relationships are not important.
53. Families are very important.
60. Family relationships cause more trouble than they're worth.
69. Family relationships can be very valuable.

Self Esteem

3. Overall, I am satisfied with myself.
26. I feel I do not have much to be proud of.
35. At time I think I´m no good at all.
44. I feel that I have many good personal qualities.
68. I wish I had more respect for myself.

Satisfaction with Personal Sexuality

7. I am not happy with my sex life.
18. I am not satisfied with my sexual behavior (sex life).
33. I´m happy with my sexual behavior now.
39. Sexually I feel like a failure.
57. I am very satisfied with my sexual activities just the way they are.

Satisfaction with Social Relationships

1. I am very happy with my friendships.
32. I don´t have enough friends.
38. My friendships are not as good as I would like them to be.
43. I wish my friendships were better.
67. I feel good having as many friends as I have.

BEHAVIOR INVENTORY

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ______________________________

Teacher's name: ______________________________________

Your birth date: Month _________ Day _____

Your sex (Check one): Male ___ Female ____

Your grade level in school (Check one): 9 ______
10 ______
11 ______
12 ______

Part 1.

The questions below ask how often you have done some things. Some of the questions are personal and ask about your social life and sex life. Some questions will not apply to you. Please do not conclude from the questions that you should have had all of the experiences the questions ask about. Instead, just mark whatever answer describes you best.

Circle: 1 = if you do it Almost Never, which means about 5% of the time or less.
2 = if you do it Sometimes, which means about 25% of the time.
3 = if you do it Half the Time, which means about 50% of the time.
4 = if you do it Usually, which means about 75% of the time.
5 = if you do it Almost Always, which means about 95% of the time or more.
DNA = if the question Does Not Apply to you.

		Almost Never	Sometimes	Half the Time	Usually	Almost Always	Does Not Apply
1.	When things you've done turn out poorly, how often do you take responsibility for your behavior and its consequences?	1	2	3	4	5	DNA
2.	When things you've done turn out poorly, how often do you blame others?	1	2	3	4	5	DNA
3.	When you are faced with a decision, how often do you take responsibility for making a decision about it?	1	2	3	4	5	DNA
4.	When you have to make a decision, how often do you think hard about the consequences of each possible choice?	1	2	3	4	5	DNA
5.	When you have to make a decision, how often do you get as much information as you can before making the decision?	1	2	3	4	5	DNA
6.	When you have to make a decision, how often do you first discuss it with others?	1	2	3	4	5	DNA
7.	When you have to make a decision about your sexual behavior (for example, going out on a date, holding hands, kissing, petting, or having sex), how often do you take responsibility for the consequences?	1	2	3	4	5	DNA
8.	When you have to make a decision about your sexual behavior, how often do you think hard about the consequences of each possible choice?	1	2	3	4	5	DNA

		Almost Never	Sometimes	Half the Time	Usually	Almost Always	Does Not Apply
9.	When you have to make a decision about your sexual behavior, how often do you first get as much information as you can?	1	2	3	4	5	DNA
10.	When you have to make a decision about your sexual behavior, how often do you first discuss it with others?	1	2	3	4	5	DNA
11.	When you have to make a decision about your sexual behavior, how often do you make it on the spot without worrying about the consequences?	1	2	3	4	5	DNA
12.	When a friend wants to talk with you, how often are you able to clear your mind and really listen to what your friend has to say?	1	2	3	4	5	DNA
13.	When a friend is talking with you, how often do you ask questions if you don′t understand what your friend is saying?	1	2	3	4	5	DNA
14.	When a friend is talking with you, how often do you nod your head and say "yes" or something else to show that you are interested?	1	2	3	4	5	DNA
15.	When you want to talk with a friend, how often are you able to get your friend to really listen to you?	1	2	3	4	5	DNA
16.	When you talk with a friend, how often do you ask for your friend′s reaction to what you′ve said?	1	2	3	4	5	DNA
17.	When you talk with a friend, how often do you let your feelings show?	1	2	3	4	5	DNA
18.	When you are with a friend you care about, how often do you let that friend know you care?	1	2	3	4	5	DNA
19.	When you talk with a friend, how often do you include statements like "my feelings are...," "the way I think is...," or "it seems to me"?	1	2	3	4	5	DNA
20.	When you are alone with a date or boy/girlfriend, how often can you tell him/her your feelings about what you want to do and do not want to do sexually? (If you are a boy, boy/girlfriend means girlfriend; if you are a girl, it means boyfriend.)	1	2	3	4	5	DNA

	Almost Never	Sometimes	Half the Time	Usually	Almost Always	Does Not Apply
21. If a boy/girl puts pressure on you to be involved sexually and you don't want to be involved, how often do you say "no"? (If you are a boy, boy/girl means girl; if you are a girl, it means boy.)	1	2	3	4	5	DNA
22. If a boy/girl puts pressure on you to be involved sexually and you don't want to be involved, how often do you succeed in stopping it?	1	2	3	4	5	DNA
23. If you have sexual intercourse with your boy/girlfriend, how often can you talk with him/her about birth control?	1	2	3	4	5	DNA
24. If you have sexual intercourse and want to use birth control, how often do you insist on using birth control?	1	2	3	4	5	DNA

Part 2.

In this section, we want to know how uncomfortable you are doing different things. Being "uncomfortable" means that it is difficult for you and it makes you nervous and up-tight. For each item, circle the number that describes you best, but if the item doesn't apply to you, circle DNA.

Circle: 1 = if you are Comfortable.
2 = if you are A Little Uncomfortable.
3 = if you are Somewhat Uncomfortable.
4 = if you are Very Uncomfortable.
DNA = if the question Does Not Apply to you.

		Comfortable	A Little Uncomfortable	Somewhat Uncomfortable	Very Uncomfortable	Does Not Apply
25.	Getting together with a group of friends of the opposite sex.	1	2	3	4	DNA
26.	Going to a party.	1	2	3	4	DNA
27.	Talking with teenagers of the opposite sex.	1	2	3	4	DNA
28.	Going out on a date.	1	2	3	4	DNA
29.	Talking with friends about sex.	1	2	3	4	DNA
30.	Talking with a date or boy/girlfriend about sex. (If you are a boy, boy/girlfriend means girlfriend; if you are a girl, it means boyfriend.)	1	2	3	4	DNA
31.	Talking with parents about sex.	1	2	3	4	DNA
32.	Talking with friends about birth control.	1	2	3	4	DNA
33.	Talking with a date or boy/girlfriend about birth control. (If you are a boy, boy/girlfriend means girlfriend; if you are a girl, it means boyfriend.)	1	2	3	4	DNA
34.	Talking with parents about birth control.	1	2	3	4	DNA
35.	Expressing concern and caring for others.	1	2	3	4	DNA
36.	Telling a date or boy/girlfriend what you want to do and do not want to do sexually.	1	2	3	4	DNA
37.	Saying "no" to a sexual come-on.	1	2	3	4	DNA
38.	Having your current sex life, whatever it may be (it may be doing nothing, kissing, petting, or having intercourse).	1	2	3	4	DNA

If you are not having sexual intercourse, circle DNA in the four questions below.

	Comfortable	A Little Uncomfortable	Somewhat Uncomfortable	Very Uncomfortable	Does Not Apply
39. Insisting on using some form of birth control, if you are having sex.	1	2	3	4	DNA
40. Buying contraceptives at a drug store, if you are having sex.	1	2	3	4	DNA
41. Going to a doctor or clinic for contraception, if you are having sex.	1	2	3	4	DNA
42. Using some form of birth control, if you are having sex.	1	2	3	4	DNA

Part 3.

Circle the correct answer to the following two questions.

43. Have you ever had sex (sexual intercourse)? yes no

44. Have you had sex (sexual intercourse) during the last month? yes no

Part 4.

The following questions ask how many times you did some things during the last month. Put a number in the right hand space to show the number of times you engaged in that activity. If you did not do that during the last month, put a "0" in the space.

Think CAREFULLY about the times that you have had sex during the last month. Think also about the number of times you did not use birth control and the number of times you used different types of birth control.

45. Last month, how many times did you have sex (sexual intercourse)? ___ times in the last month

46. Last month, how many times did you have sex when you or your partner did not use any form of birth control? ___ times in the last month

47. Last month, how many times did you have sex when you or your partner used a diaphragm, withdrawal (pulling out before releasing fluid), rhythm (not having sex on fertile days), or foam without condoms? ___ times in the last month

48. Last month, how many times did you have sex when you or your partner used the pill, condoms (rubbers), or an IUD? ___ times in the last month

(If you add your answers to questions #46, #47, and #48, the total should equal your answer to #45. If it does not, please correct your answers.)

49. During the last month, how many times have you had a conversation or discussion about sex with your parents? ___ times in the last month

50. During the last month, how many times have you had a conversation or discussion about sex with your friends? ___ times in the last month

51. During the last month, how many times have you had a conversation or discussion about sex with a date or boy/girlfriend? (If you are a boy, boy/girlfriend means girlfriend; if you are a girl, it means boyfriend.) ___ times in the last month

52. During the last month, how many times have you had a conversation or discussion about birth control with your parents? ___ times in the last month

53. During the last month, how many times have you had a conversation or discussion about birth control with your friends? ___ times in the last month

54.	During the last month, how many times have you had a conversation or discussion about birth control with a date or boy/girlfriend?	___ times in the last month

Thank you for completing the questionnaire.

KNOWLEDGE, ATTITUDE, AND BEHAVIOR QUESTIONNAIRE

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ____

Teacher's name: ____

Your birth date: Month ________ Day _____________

Your sex (Check one): Male _______ Female _______

Your grade level in school (Check one):
9 ________
10 ________
11 ________
12 ________

Part 1.

Circle the one best answer to each of the questions below.

1. Some contraceptives:

   a. can be obtained only with a doctor's prescription.
   b. are available at family planning clinics.
   c. can be bought over the counter at drug stores.
   d. can be obtained by people under 18 without their parents' permission.
   e. all of the above.

2. If 10 couples have sexual intercourse regularly without using any kind of birth control, the number of couples who become pregnant by the end of 1 year is about:

   a. one.
   b. three.
   c. six.
   d. nine.
   e. none of the above.

3. People having sexual intercourse can best prevent getting a sexually transmitted disease (VD or STD) by using:

   a. condoms (rubbers).
   b. contraceptive foam.
   c. the pill.
   d. withdrawal (pulling out).

4. If a couple has sexual intercourse and uses no birth control, the woman might get pregnant:

   a. any time during the month.
   b. only 1 week before menstruation begins.
   c. only during menstruation.
   d. only 1 week after menstruation begins.
   e. only 2 weeks after menstruation begins.

5. The method of birth control which is least effective is:

   a. a condom with foam.
   b. the diaphragm with spermicidal jelly.
   c. withdrawal (pulling out).
   d. the pill.
   e. abstinence (not having intercourse).

6. It is possible for a woman to become pregnant:

   a. the first time she has sexual intercourse.
   b. if she has sexual intercourse during her menstrual period.
   c. if she has sexual intercourse standing up.
   d. if sperm get near the opening of the vagina, even though the man's penis does not enter her body.
   e. all of the above.

7. In general, children born to young teenage parents:

   a. have few problems because their parents are emotionally mature.
   b. have a greater chance of being abused by their parents.
   c. have normal birth weight.
   d. have a greater chance of being healthy.
   e. none of the above.

8. If people have sexual intercourse, the advantage of using condoms is that they:

   a. help prevent getting or giving VD.
   b. can be bought in drug stores by either sex.
   c. do not have dangerous side effects.
   d. do not require a prescription.
   e. all of the above.

9. Most unmarried girls who have children while still in high school:

   a. depend upon their parents for support.
   b. finish high school and graduate with their class.
   c. never have to be on public welfare.
   d. have the same social lives as their peers.
   e. all of the above.

10. People choosing a birth control method:

   a. should think only about the cost of the method.
   b. should choose whatever method their friends are using.
   c. should learn about all the methods before choosing the one that´s best for them.
   d. should get the method that´s easiest to get.
   e. all of the above.

Part 2.

This part is NOT a knowledge test. We are interested in what you believe about some important issues. Please rate each statement according to how much you agree or disagree with it. Everyone will have different answers. Your answer is correct if it describes you very well.

Circle: 1 = if you Strongly Disagree with the statement.
2 = if you Somewhat Disagree with the statement.
3 = if you feel Neutral about the statement.
4 = if you Somewhat Agree with the statement.
5 = if you Strongly Agree with the statement.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
11. Unmarried people should not have sex.	1	2	3	4	5
12. I have my own set of rules to guide my sexual behavior (sex life).	1	2	3	4	5
13. Birth control is not very important.	1	2	3	4	5
14. People should not have sex before marriage.	1	2	3	4	5
15. I know for sure what is right and wrong sexually for me.	1	2	3	4	5
16. Birth control is not as important as some people say.	1	2	3	4	5
17. I have trouble knowing what my values are about my personal sexual behavior.	1	2	3	4	5
18. More people should be aware of the importance of birth control.	1	2	3	4	5
19. People should have sex only if they are married.	1	2	3	4	5
20. I´m confused about my personal sexual values and beliefs.	1	2	3	4	5
21. Two people having sex should use some form of birth control if they aren´t ready for a child.	1	2	3	4	5
22. It is all right for two people to have sex before marriage if they are in love.	1	2	3	4	5
23. I´m confused about what I should and should not do sexually.	1	2	3	4	5
24. If two people have sex and aren´t ready to have a baby, it is very important that they use birth control.	1	2	3	4	5
25. It is all right for two people to have sex before marriage.	1	2	3	4	5

Part 3.

The following parts ask questions that are personal and ask about your social life and sex life. Some questions will not apply to you. Please do not conclude from these questions that you should have had all of the experiences the questions ask about. Instead, just mark whatever answer describes you best.

In this section, we want to know how uncomfortable you are doing different things. Being "uncomfortable" means that it is difficult for you and you feel nervous and uptight.

Circle: 1 = if you are Comfortable.
2 = if you are A Little Uncomfortable.
3 = if you are Somewhat Uncomfortable.
4 = if you are Very Uncomfortable.
DNA = if the question Does Not Apply to you.

	Comfortable	A Little Uncomfortable	Somewhat Uncomfortable	Very Uncomfortable	Does Not Apply
26. Talking with friends about sex.	1	2	3	4	DNA
27. Talking with your boy/girlfriend about sex. ("boy/girlfriend" means "boyfriend" if you are a girl, and it means "girlfriend" if you are a boy.)	1	2	3	4	DNA
28. Talking with parents about sex.	1	2	3	4	DNA
29. Talking with friends about birth control.	1	2	3	4	DNA
30. Talking with your boy/girlfriend about birth control.	1	2	3	4	DNA
31. Talking with parents about birth control.	1	2	3	4	DNA
32. Having your current sex life, whatever it may be (it may be doing nothing, kissing, petting, or having intercourse).	1	2	3	4	DNA

If you are not having sexual intercourse, circle DNA in the three questions below.

33. Buying contraceptives at a drug store, if you are having sex.	1	2	3	4	DNA
34. Going to a doctor or clinic for contraception, if you are having sex.	1	2	3	4	DNA
35. Using birth control, if you are having sex.	1	2	3	4	DNA

Part 4.

The questions below ask how often you do some things.

Circle: 1 = if you do it Almost Never, which means about 5% of the time or less.
2 = if you do it Sometimes, which means about 25% of the time.
3 = if you do it Half the Time, which means about 50% of the time.
4 = if you do it Usually, which means about 75% of the time.
5 = if you do it Almost Always, which means about 95% of the time or more.
DNA = if the question Does Not Apply to you.

	Almost Never	Sometimes	Half the Time	Usually	Almost Always	Does Not Apply
36. When you have to make a decision about your sexual behavior (holding hands, kissing, petting, or having sex), how often do you think hard about the consequences of each possible alternative?	1	2	3	4	5	DNA
37. When you have to make a decision about your sexual behavior, how often do you first get as much information as you can?	1	2	3	4	5	DNA
38. When you have to make a decision about your sexual behavior, how often do you first discuss it with other people?	1	2	3	4	5	DNA
39. When you have to make a decision about your sexual behavior, how often do you make it on the spot without thinking about the consequences?	1	2	3	4	5	DNA
40. If a boy/girl puts pressure on you to be involved sexually and you don´t want to be involved, how often do you stop him/her?	1	2	3	4	5	DNA
41. If you have sexual intercourse with your boy/girlfriend, how often can you talk with him/her about using birth control?	1	2	3	4	5	DNA

Part 5.

Circle the correct answer to the following two questions.

42. Have you ever had sexual intercourse?	yes	no
43. Have you had sexual intercourse during the last month?	yes	no

Part 6.

The following questions ask about activities during the last month. Put a number in the right hand space which shows the number of times you engaged in that activity. Put a "0" in that space if you did not engage in that activity during the last month.

Think CAREFULLY about the times that you have had sex during the last month. Think also about the number of times you did not use birth control and the number of times you used different types of birth control.

44. Last month, how many times did you have sexual intercourse? ___ times in the last month

45. Last month, how many times did you have sex when you or your partner did not use any form of birth control? ___ times in the last month

46. Last month, how many times did you have sex when you or your partner used a diaphragm, withdrawal (pulling out before releasing fluid), rhythm (not having sex on fertile days), or foam without condoms? ___ times in the last month

47. Last month, how many times did you have sex when you or your partner used the pill, condoms (rubbers), or an IUD? ___ times in the last month

(If you add your answers to questions #45, #46, and #47, the total should equal your answer to #44. If it does not, please correct your answers.)

48. During the last month, how many times have you had a conversation or discussion about sex with your parents? ___ times in the last month

49. During the last month, how many times have you had a conversation or discussion about sex with your friends? ___ times in the last month

50. During the last month, how many times have you had a conversation or discussion about sex with a date or boy/girlfriend? (If you are a boy, boy/girlfriend means girlfriend; if you are a girl, it means boyfriend.) ___ times in the last month

51. During the last month, how many times have you had a conversation or discussion about birth control with your parents? ___ times in the last month

52. During the last month, how many times have you had a conversation or discussion about birth control with your friends? ___ times in the last month

53. During the last month, how many times have you had a conversation or discussion about birth control with a date or boy/girlfriend? ___ times in the last month

Thank you for completing the questionnaire.

COURSE EVALUATION

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ______________________________

Teacher's name: ______________________________________

Your birth date: Month __________ Day ______

Your sex (Check one): Male ___ Female ____

Your grade level in school (Check one): 9 ______
10 ______
11 ______
12 ______

__Part 1__.

Below is a list of questions about your teacher. Now that this class is over, please answer each question by circling one number based upon this 5-point scale:

1 = Not at All
2 = A Small Amount
3 = A Medium Amount
4 = A Large Amount
5 = A Great Deal

		Not at All	A Small Amount	A Medium Amount	A Large Amount	A Great Deal
1.	Was the teacher enthusiastic about teaching this course?	1	2	3	4	5
2.	Was the teacher uncomfortable discussing different things about sex?	1	2	3	4	5
3.	Did the teacher discuss topics in a way that made students feel uncomfortable?	1	2	3	4	5
4.	Did the teacher talk at a level that the students could understand?	1	2	3	4	5
5.	Did the teacher care about the students?	1	2	3	4	5
6.	Did the teacher show respect toward the students?	1	2	3	4	5
7.	Did the students trust the teacher?	1	2	3	4	5
8.	Did the teacher get along with the students?	1	2	3	4	5
9.	Did the teacher encourage students to talk about their feelings and opinions?	1	2	3	4	5
10.	Did the teacher talk too much about what's right and wrong?	1	2	3	4	5
11.	Did the teacher listen carefully to the students?	1	2	3	4	5
12.	Did the teacher discourage students from hurting others in sexual situations (such as knowingly spreading VD or forcing someone to have sex)?	1	2	3	4	5
13.	Did the teacher encourage students to think about the consequences before having sexual relations?	1	2	3	4	5
14.	Did the teacher encourage students to think about their own values about sexuality?	1	2	3	4	5
15.	Did the teacher encourage the use of birth control to avoid an unwanted pregnancy?	1	2	3	4	5
16.	Did the teacher encourage students to talk with their parents about sexuality?	1	2	3	4	5

Part 2.

Below is a list of questions about you and the course. Continue to answer each question by circling one number based upon the same 5-point scale:

1 = Not at All
2 = A Small Amount
3 = A Medium Amount
4 = A Large Amount
5 = A Great Deal

		Not at All	A Small Amount	A Medium Amount	A Large Amount	A Great Deal
17.	Were you bored by the course?	1	2	3	4	5
18.	Did students participate in class discussions?	1	2	3	4	5
19.	Were you encouraged to ask any questions you had about sex?	1	2	3	4	5
20.	Was it hard for you to talk about your own thoughts and feelings?	1	2	3	4	5
21.	Was it hard for you to ask questions and talk about sexual topics?	1	2	3	4	5
22.	Did you show concern for the other students in the class?	1	2	3	4	5
23.	Did the other students show concern for you?	1	2	3	4	5
24.	Were students' opinions kept confidential (not spread outside the classroom)?	1	2	3	4	5
25.	Were you permitted to have values or opinions that were different from others in the class?	1	2	3	4	5

Part 3.

These five questions should be answered using another 5-point scale. Circle the number that best describes your opinion, but if you don't know, circle DK.

1 = Very Poor
2 = Poor
3 = Average
4 = Good
5 = Excellent
DK = Don't Know

		Very Poor	Poor	Average	Good	Excellent	Don't Know
26.	What is your opinion of the teacher?	1	2	3	4	5	DK
27.	What is your opinion of the topics covered in the course?	1	2	3	4	5	DK
28.	What is your opinion of the materials used, such as books and films?	1	2	3	4	5	DK
29.	What is your opinion of the organization and format of the program, such as length, location, and time?	1	2	3	4	5	DK
30.	What is your opinion of the overall program?	1	2	3	4	5	DK

31. What things about the program did you particularly like?

32. What things about the program do you think should be changed? How?

Thank you for completing the questionnaire.

ASSESSMENT OF COURSE IMPACT

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ______________________________

Teacher's name: ______________________________________

Your birth date: Month _________ Day _____

Your sex (Check one): Male ___ Female ____

Your grade level in school (Check one):
9 ______
10 ______
11 ______
12 ______

Part 1.

Directions: Now that this sexuality education course is over, we would like to know how it may have changed you, if at all. Please answer each question by circling the number that best describes how you have changed because of this course.

Circle: 1 = Much Less
2 = Somewhat Less
3 = About the Same
4 = Somewhat More
5 = Much More

		Much Less	Somewhat Less	About the Same	Somewhat More	Much More
1.	Do you know less or more about sexuality?	1	2	3	4	5
2.	Do you understanding yourself and your behavior less or more?	1	2	3	4	5
3.	Are your attitudes and values about your own sexual behavior less or more clear?	1	2	3	4	5
4.	Do you now feel that using birth control when people are not ready to have children is less or more important?	1	2	3	4	5
5.	Do you talk about sexuality (going out, having sex, birth control, or male and female sex roles) with your friends less or more?	1	2	3	4	5
6.	Do you talk about sexuality with your boy/girlfriend less or more?	1	2	3	4	5
7.	Do you talk about sexuality with your parents less or more?	1	2	3	4	5
8.	When you talk about sexuality with others (such as your friends, boy/girlfriend, and parents) are you less or more comfortable?	1	2	3	4	5
9.	Do you talk about sexuality less or more effectively (that is, are you less or more able to talk about your thoughts, feelings, and needs and to listen carefully)?	1	2	3	4	5
10.	Are you less or more likely to have sex?	1	2	3	4	5
11.	If you have sex, would you be less or more likely to use birth control?	1	2	3	4	5
12.	If you have sex, would you be less or more comfortable using birth control?	1	2	3	4	5
13.	Do you respect yourself less or more?	1	2	3	4	5

	Much Less	Somewhat Less	About the Same	Somewhat More	Much More
14. Are you less or more satisfied with your social life?	1	2	3	4	5
15. Are you less or more satisfied with your sex life whatever it may be (it may be doing nothing, kissing, petting, or having sex)?	1	2	3	4	5

Part 2.

We are still interested in knowing about any ways you may have changed because of this course. Please answer the following questions by circling the number that describes you best:

1 = Much Worse
2 = Somewhat Worse
3 = About the Same
4 = Somewhat Better
5 = Much Better

	Much Worse	Somewhat Worse	About the Same	Somewhat Better	Much Better
16. Do you now make worse or better decisions about your social life?	1	2	3	4	5
17. Do you now make worse or better decisions about your physical sexual behavior?	1	2	3	4	5
18. Do you now get along with your friends worse or better?	1	2	3	4	5

Thank you for completing the questionnaire.

COURSE ASSESSMENT FOR PARENTS

We are trying to find out if this program is successful. You can help us by completing this questionnaire.

To keep your answers confidential and private, do NOT put your name anywhere on this questionnaire. Please use a regular pen or pencil so that all questionnaires will look about the same and no one will know which is yours.

Because this study is important, your answers are also important. Please answer each question carefully.

Thank you for your help.

Name of school or organization
where course was taken: ____

Teacher's name: ____

Your birth date: Month ________ Day ____________

Your sex (Check one): Male _______ Female _______

Your grade level in school (Check one): 9 ________
10 ________
11 ________
12 ________

Now that your teenager´s sex education course is over, we are interested in your ideas about whether it changed him or her. For each question, please circle the number that best describes your opinion. If you don´t know, circle DK.

1 = Much Less
2 = Somewhat Less
3 = About the Same
4 = Somewhat More
5 = Much More
DK = Don´t Know

	Much Less	Somewhat Less	About the Same	Somewhat More	Much More	Don't Know
1. Does your teenager know less or more about sexuality?	1	2	3	4	5	DK
2. Are your teenager´s attitudes and values about sexuality less or more clear?	1	2	3	4	5	DK
3. Are you less or more comfortable talking about sexuality with your teenager?	1	2	3	4	5	DK
4. Have you actually talked about sexuality with your teenager less or more?	1	2	3	4	5	DK
5. Does your teenager talk and listen to you about sexuality less or more effectively? That is, is your teenager less or more able to talk about thoughts, feelings, and needs, and to listen carefully?	1	2	3	4	5	DK
6. Is your teenager less or more likely to make good decisions about social and sexual behavior? That is, is your teenager less or more able to examine alternatives and consider consequences?	1	2	3	4	5	DK
7. Is your teenager less or more likely to have sex soon because of this course?	1	2	3	4	5	DK

These five questions should be answered using another 5-point scale. Again, if you don´t know, circle DK.

1 = Very Poor
2 = Poor
3 = Average
4 = Good
5 = Excellent
DK = Don´t Know

	Very Poor	Poor	Average	Good	Excellent	Don't Know
7. What is your opinion of the teacher?	1	2	3	4	5	DK
8. What is your opinion of the topics covered in the course?	1	2	3	4	5	DK
9. What is your opinion of the materials used, such as books and films?	1	2	3	4	5	DK
10. What is your opinion of the organization and format of the program, such as length, location, and time?	1	2	3	4	5	DK
11. What is your opinion of the overall program?	1	2	3	4	5	DK

12. What things about the program did you particularly like?

13. What things about the program do you think should be changed? How?

Thank you for completing the questionnaire.